

【書類名】 特許願

【整理番号】 0290541104

【提出日】 平成14年11月13日

【あて先】 特許庁長官殿

【国際特許分類】 G06F 17/30

【発明者】

【住所又は居所】 東京都品川区北品川 6 丁目 7 番 3 5 号 ソニー株式会社
内

【氏名】 大野 潮満

【特許出願人】

【識別番号】 000002185

【氏名又は名称】 ソニー株式会社

【代理人】

【識別番号】 100082131

【弁理士】

【氏名又は名称】 稲本 義雄

【電話番号】 03-3369-6479

【手数料の表示】

【予納台帳番号】 032089

【納付金額】 21,000円

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【包括委任状番号】 9708842

【プルーフの要否】 要

日本国特許庁
JAPAN PATENT OFFICE

SO 391311
US

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日 2002年11月13日
Date of Application:

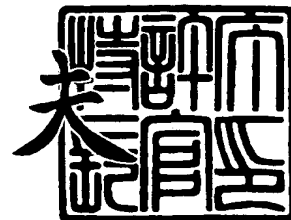
出願番号 特願2002-329492
Application Number:
[ST. 10/C]: [JP 2002-329492]

出願人 ソニー株式会社
Applicant(s):

2003年 8月11日

特許庁長官
Commissioner,
Japan Patent Office

今井 康



出証番号 出証特2003-3064374

【書類名】 明細書

【発明の名称】 情報処理装置および方法、記録媒体、並びにプログラム

【特許請求の範囲】

【請求項 1】 サイトを構成するページのデータを取得する取得手段と、
前記取得手段により取得された前記ページのデータを用いて、前記ページ内に
出現する単語を抽出する抽出手段と、

前記抽出手段により抽出された前記単語が、前記ページ内で出現する回数をカ
ウントするカウント手段と、

前記取得手段で取得されたページ間のリンク構造を解析し、前記カウント手段
によるカウントの値を用いて、リンク関係にある前記ページ間の第 1 の重みを生
成する第 1 の生成手段と、

前記第 1 の生成手段により生成された前記第 1 の重みを用いて、所定のページ
とリンク関係にある他のページ同士のページ間の第 2 の重みを生成する第 2 の生
成手段と、

前記第 2 の生成手段により生成された前記第 2 の重みを用いて、S D F データ
または C D F データの少なくとも一方のデータを生成する第 3 の生成手段と、

前記第 3 の生成手段により生成された前記データを用いて、I S D F または I
C D F の少なくとも一方に基づくページモデル拡張処理により所定の値を算出す
る算出手段と

を含むことを特徴とする情報処理装置。

【請求項 2】 前記算出手段により算出された前記所定の値を用いて、前記
取得されたページ間の所定のページ間の関連度を算出する第 2 の算出手段を

さらに含むことを特徴とする請求項 1 に記載の情報処理装置。

【請求項 3】 前記第 2 の生成手段が、前記所定のページをリンク元とし、
そのリンク元からリンクが張られているリンク先のページ同士のページ間の前記
第 2 の重みを算出する場合、前記第 3 の生成手段は、前記 S D F データを生成し
、前記算出手段は、前記 I S D F に基づくページモデル拡張処理により前記所定
の値を算出し、

前記第 2 の生成手段が、前記所定のページをリンク先とし、そのリンク先にリ

リンクを張っているリンク元のページ同士のページ間の前記第2の重みを算出する場合、前記第3の生成手段は、前記CDFデータを生成し、前記算出手段は、前記ICDFに基づくページモデル拡張処理により前記所定の値を算出し、

前記第2の生成手段が、前記所定のページをリンク元とし、そのリンク元からリンクが張られているリンク先のページ同士のページ間の前記第2の重みと、前記所定のページをリンク先とし、そのリンク先にリンクを張っているリンク元のページ同士のページ間の前記第2の重みをそれぞれ算出する場合、前記第3の生成手段は、前記SDFデータと前記CDFデータをそれぞれ生成し、前記算出手段は、前記ISDFと前記ICDFに基づくページモデル拡張処理により前記所定の値を算出する

ことを特徴とする請求項1に記載の情報処理装置。

【請求項4】 前記算出手段は、前記所定のページ内における所定の単語の出現回数と、前記所定のページと前記第2の生成手段で生成されたリンク関係にあるページのうち、前記所定の単語を含むページに対応する前記第3の生成手段により生成された前記データを用いた演算により、前記所定の値を算出する

ことを特徴とする請求項1に記載の情報処理装置。

【請求項5】 前記第2の算出手段により算出された前記関連度を記憶する記憶手段と、

所定のページに関連があるページの情報の提供が要求された場合、前記記憶手段に記憶されている前記関連度を参照して、前記所定のページと関連度が高いページの情報を提供する提供手段と

をさらに含むことを特徴とする請求項1に記載の情報処理装置。

【請求項6】 前記提供手段は、前記情報を提供する際、前記所定のページと関連する広告に関する情報も提供する

ことを特徴とする請求項5に記載の情報処理装置。

【請求項7】 サイトを構成するページのデータを取得する取得ステップと、

前記取得ステップの処理で取得された前記ページのデータを用いて、前記ページ内に出現する単語を抽出する抽出ステップと、

前記抽出ステップの処理で抽出された前記単語が、前記ページ内で出現する回数をカウントするカウントステップと、

前記取得ステップの処理で取得されたページ間のリンク構造を解析し、前記カウントステップの処理によるカウントの値を用いて、リンク関係にある前記ページ間の第1の重みを生成する第1の生成ステップと、

前記第1の生成ステップの処理で生成された前記第1の重みを用いて、所定のページとリンク関係にある他のページ同士のページ間の第2の重みを生成する第2の生成ステップと、

前記第2の生成ステップの処理で生成された前記第2の重みを用いて、SDFデータまたはCDFデータの少なくとも一方のデータを生成する第3の生成ステップと、

前記第3の生成ステップの処理で生成された前記データを用いて、ISDFまたはICDFの少なくとも一方に基づくページモデル拡張処理により所定の値を算出する第1の算出ステップと、

前記第1の算出ステップの処理で算出された前記所定の値を用いて、取得されたページ内の所定のページ間の関連度を算出する第2の算出ステップと

を含むことを特徴とする情報処理方法。

【請求項8】 サイトを構成するページのデータを取得する取得ステップと

、
前記取得ステップの処理で取得された前記ページのデータを用いて、前記ページ内に出現する単語を抽出する抽出ステップと、

前記抽出ステップの処理で抽出された前記単語が、前記ページ内で出現する回数をカウントするカウントステップと、

前記取得ステップの処理で取得されたページ間のリンク構造を解析し、前記カウントステップの処理によるカウントの値を用いて、リンク関係にある前記ページ間の第1の重みを生成する第1の生成ステップと、

前記第1の生成ステップの処理で生成された前記第1の重みを用いて、所定のページとリンク関係にある他のページ同士のページ間の第2の重みを生成する第2の生成ステップと、

前記第 2 の生成ステップの処理で生成された前記第 2 の重みを用いて、S D F データまたは C D F データの少なくとも一方のデータを生成する第 3 の生成ステップと、

前記第 3 の生成ステップの処理で生成された前記データを用いて、I S D F または I C D F の少なくとも一方に基づくページモデル拡張処理により所定の値を算出する第 1 の算出ステップと、

前記第 1 の算出ステップの処理で算出された前記所定の値を用いて、取得されたページ内の所定のページ間の関連度を算出する第 2 の算出ステップと

を含むことを特徴とするコンピュータが読み取り可能なプログラムが記録されている記録媒体。

【請求項 9】 サイトを構成するページのデータを取得する取得ステップと

、
前記取得ステップの処理で取得された前記ページのデータを用いて、前記ページ内に出現する単語を抽出する抽出ステップと、

前記抽出ステップの処理で抽出された前記単語が、前記ページ内で出現する回数をカウントするカウントステップと、

前記取得ステップの処理で取得されたページ間のリンク構造を解析し、前記カウントステップの処理によるカウントの値を用いて、リンク関係にある前記ページ間の第 1 の重みを生成する第 1 の生成ステップと、

前記第 1 の生成ステップの処理で生成された前記第 1 の重みを用いて、所定のページとリンク関係にある他のページ同士のページ間の第 2 の重みを生成する第 2 の生成ステップと、

前記第 2 の生成ステップの処理で生成された前記第 2 の重みを用いて、S D F データまたは C D F データの少なくとも一方のデータを生成する第 3 の生成ステップと、

前記第 3 の生成ステップの処理で生成された前記データを用いて、I S D F または I C D F の少なくとも一方に基づくページモデル拡張処理により所定の値を算出する第 1 の算出ステップと、

前記第 1 の算出ステップの処理で算出された前記所定の値を用いて、取得され

たページ内の所定のページ間の関連度を算出する第 2 の算出ステップと
をコンピュータに実行させることを特徴とするプログラム。

【発明の詳細な説明】

【 0 0 0 1 】

【発明の属する技術分野】

本発明は情報処理装置および方法、記録媒体、並びにプログラムに関し、特に、ネットワーク上で開設されているホームページなどの検索に用いて好適な情報処理装置および方法、記録媒体、並びにプログラムに関する。

【 0 0 0 2 】

【従来の技術】

近年、インターネットの普及により、そのインターネット上で開設されているホームページの数も増大しつつある。それらのホームページは、企業だけでなく、個人ユーザも開設しているため、その数は、膨大なものとなっている。それら膨大な数のホームページから、ユーザが所望の情報を掲載したホームページを探し出すということは大変な手間がかかることであった。

【 0 0 0 3 】

そのような手間を省くために、キーワードなどを入力するだけで、所望のホームページが検索できるような、俗に検索エンジンなどと称されるホームページ、例えば、Y a h o o (商標)、g o o (商標)、E x c i t e (商標)、G o o g l e (商標)、N e t s c a p e (商標)がサービスの提供を開始している。

【 0 0 0 4 】

これらの検索エンジンは、ユーザが入力したキーワードを含み、キーワードの特徴に近い類似したホームページを探す際に適しているが、その検索結果以外にもユーザが所望するページが多い。

【 0 0 0 5 】

そのため、幾つかの検索エンジンでは、関連ページ検索などと称される関連ページ検索エンジンのサービスを開始している。例えば、特許文献 1 や、G o o g l e の検索結果の各々ページに対する関連ページ検索、G o o g l e T o o l b a r の関連ページ検索ボタン、N e t s c a p e N a v i g a t o r などの

ブラウザに表示される関連サイト検索ボタンなどがある。

【0006】

【特許文献1】

特開 2002-149698 号公報（第4—7頁）

【0007】

【発明が解決しようとする課題】

関連ページ検索エンジンを用いた検索は、ユーザが閲覧中のページ、あるいは検索エンジンの検索結果の所定のページに対して関連するページが検索される。その検索は、WWW（World Wide Web）のリンク構造を考慮するものもあったが、関連ページの検索が必ずしも精度良く行われているとは限らなかった。

【0008】

これは、従来のページの特徴抽出によるページモデルの生成は、関連ページ検索ではなく、検索エンジン、つまり、入力されるキーワードや自然言語と検索対象となるページとの類似度を求めるための手段であったため、関連ページを検索する場合のページの特徴抽出には適していなかったためである。関連ページ検索では関連ページ検索に適したページの特徴抽出に基づくページモデルの生成が必要である。

【0009】

本発明はこのような状況に鑑みてなされたものであり、リンク構造のうち、兄弟関係(Sibling関係)、あるいは共通親関係(Co-Parent関係)、またはその両方を考慮したページの特徴抽出により関連ページ検索に適したページモデルを生成し、このページモデルに基づく関連ページ検索エンジンを提供することにより、関連ページの検索をより精度良く行われるようにすることを目的とする。

【0010】

【課題を解決するための手段】

本発明の情報処理装置は、サイトを構成するページのデータを取得する取得手段と、取得手段により取得されたページのデータを用いて、ページ内に出現する単語を抽出する抽出手段と、抽出手段により抽出された単語が、ページ内で出現する回数をカウントするカウント手段と、取得されたページ間のリンク構造を解

析し、カウント手段によるカウントの値を用いて、リンク関係にあるページ間の第1の重みを生成する第1の生成手段と、第1の生成手段により生成された第1の重みを用いて、所定のページとリンク関係にある他のページ同士のページ間の第2の重みを生成する第2の生成手段と、第2の生成手段により生成された第2の重みを用いて、SDF (Sibling Document Frequencyの略) データまたはCDF (Co-Parent Document Frequencyの略) データの少なくとも一方のデータを生成する第3の生成手段と、第3の生成手段により生成されたデータを用いて、ISDF (Inverse Sibling Document Frequencyの略) またはICDF (Inverse Co-Parent Document Frequencyの略) の少なくとも一方に基づくページモデル拡張処理により所定の値を算出する算出手段を含むことを特徴とする。

【0011】

前記算出手段により算出された所定の値を用いて、取得されたページ内の所定のページ間の関連度を算出する第2の算出手段をさらに含むことを特徴とする。

【0012】

前記第2の生成手段が、所定のページをリンク元とし、そのリンク元からリンクが張られているリンク先のページ同士のページ間の第2の重みを算出する場合、前記第3の生成手段は、SDFデータを生成し、前記算出手段は、ISDFに基づくページモデル拡張処理により所定の値を算出し、前記第2の生成手段が、所定のページをリンク先とし、そのリンク先にリンクを張っているリンク元のページ同士のページ間の第2の重みを算出する場合、前記第3の生成手段は、CDFデータを生成し、前記算出手段は、ICDFに基づくページモデル拡張処理により所定の値を算出し、前記第2の生成手段が、所定のページをリンク元とし、そのリンク元からリンクが張られているリンク先のページ同士のページ間の第2の重みと、所定のページをリンク先とし、そのリンク先にリンクを張っているリンク元のページ同士のページ間の第2の重みをそれぞれ算出する場合、前記第3の生成手段は、SDFデータとCDFデータをそれぞれ生成し、前記算出手段は、ISDFとICDFに基づくページモデル拡張処理により所定の値を算出することができる。

【0013】

前記算出手段は、所定のページ内における所定の単語の出現回数と、所定のページと前記第2の生成手段で生成されたリンク関係にあるページのうち、所定の単語を含むページに対応する前記第3の生成手段により生成されたデータを用いた演算により、所定の値を算出するようにすることができる。

【0014】

前記第2の算出手段により算出された関連度を記憶する記憶手段と、所定のページに関連があるページの情報の提供が要求された場合、前記記憶手段に記憶されている関連度を参照して、所定のページと関連度が高いページの情報を提供する提供手段とをさらに含むようにすることができる。

【0015】

前記提供手段は、情報を提供する際、所定のページと関連する広告に関する情報も提供するようにすることができる。

【0016】

本発明の情報処理方法は、サイトを構成するページのデータを取得する取得ステップと、取得ステップの処理で取得されたページのデータを用いて、ページ内に出現する単語を抽出する抽出ステップと、抽出ステップの処理で抽出された単語が、ページ内で出現する回数をカウントするカウントステップと、取得されたページ間のリンク構造を解析し、カウントステップの処理によるカウントの値を用いて、リンク関係にあるページ間の第1の重みを生成する第1の生成ステップと、第1の生成ステップの処理で生成された第1の重みを用いて、所定のページとリンク関係にある他のページ同士のページ間の第2の重みを生成する第2の生成ステップと、第2の生成ステップの処理で生成された第2の重みを用いて、SDFデータまたはCDFデータの少なくとも一方のデータを生成する第3の生成ステップと、第3の生成ステップの処理で生成されたデータを用いて、ISDFまたはICDFの少なくとも一方に基づくページモデル拡張処理により所定の値を算出する第1の算出ステップと、第1の算出ステップの処理で算出された所定の値を用いて、取得されたページ内の所定のページ間の関連度を算出する第2の算出ステップとを含むことを特徴とする。

【0017】

本発明の記録媒体のプログラムは、サイトを構成するページのデータを取得する取得ステップと、取得ステップの処理で取得されたページのデータを用いて、ページ内に出現する単語を抽出する抽出ステップと、抽出ステップの処理で抽出された単語が、ページ内で出現する回数をカウントするカウントステップと、取得されたページ間のリンク構造を解析し、カウントステップの処理によるカウントの値を用いて、リンク関係にあるページ間の第1の重みを生成する第1の生成ステップと、第1の生成ステップの処理で生成された第1の重みを用いて、所定のページとリンク関係にある他のページ同士のページ間の第2の重みを生成する第2の生成ステップと、第2の生成ステップの処理で生成された第2の重みを用いて、SDFデータまたはCDFデータの少なくとも一方のデータを生成する第3の生成ステップと、第3の生成ステップの処理で生成されたデータを用いて、ISDFまたはICDFの少なくとも一方に基づくページモデル拡張処理により所定の値を算出する第1の算出ステップと、第1の算出ステップの処理で算出された所定の値を用いて、取得されたページ内の所定のページ間の関連度を算出する第2の算出ステップとを含むことを特徴とする。

【0018】

本発明のプログラムは、サイトを構成するページのデータを取得する取得ステップと、取得ステップの処理で取得されたページのデータを用いて、ページ内に出現する単語を抽出する抽出ステップと、抽出ステップの処理で抽出された単語が、ページ内で出現する回数をカウントするカウントステップと、取得されたページ間のリンク構造を解析し、カウントステップの処理によるカウントの値を用いて、リンク関係にあるページ間の第1の重みを生成する第1の生成ステップと、第1の生成ステップの処理で生成された第1の重みを用いて、所定のページとリンク関係にある他のページ同士のページ間の第2の重みを生成する第2の生成ステップと、第2の生成ステップの処理で生成された第2の重みを用いて、SDFデータまたはCDFデータの少なくとも一方のデータを生成する第3の生成ステップと、第3の生成ステップの処理で生成されたデータを用いて、ISDFまたはICDFの少なくとも一方に基づくページモデル拡張処理により所定の値を算出する第1の算出ステップと、第1の算出ステップの処理で算出された所定の

値を用いて、取得されたページ内の所定のページ間の関連度を算出する第2の算出ステップとをコンピュータに実行させることを特徴とする。

【0019】

本発明の情報処理装置および方法、並びにプログラムにおいては、ISDFまたはICDFの少なくとも一方に基づくページモデルにより、より精度の高い関連ページ検索が行われる。

【0020】

【発明の実施の形態】

以下に、本発明の実施の形態について図面を参照して説明する。図1は、本発明の情報処理装置を含む情報処理システムの一実施の形態の構成を示す図である。ネットワーク1は、インターネットやLAN (Local Area Network) から構成されるネットワークである。ネットワーク1には、WWWサーバ2-1乃至2-3、端末3-1乃至3-3、および、検索サーバ4が接続され、相互にデータの授受を行えるように構成されている。

【0021】

以下の説明において、WWWサーバ2-1乃至2-3を個々に区別する必要がない場合、単にWWWサーバ2と記述する。他の装置に関しても同様に記述する。なお、図1には、説明の都合上、WWWサーバ2や端末3は、それぞれ3台、検索サーバ4は1台しか図示していないが、それらの装置は、ネットワーク1に複数接続されている。

【0022】

WWWサーバ2は、インターネット上のサービスの1つとして提供されているホームページを管理し、提供するサーバである。端末3は、ユーザ側の端末であり、WWWサーバ2から提供されるホームページを閲覧する機能を有する。検索サーバ4は、端末3のユーザが、WWWサーバ2で提供されるホームページに関連するページなどを検索したいときに接続されるサーバであり、ユーザの要求に対応する情報を検索し、その結果を提供する機能を有する。

【0023】

図2は、WWWサーバ2の内部構成例を示す図である。WWWサーバ2は、パーソナ

ルコンピュータなどで構成することが可能であり、そのCPU（Central Processing Unit）11は、ROM（Read Only Memory）12に記憶されているプログラムに従って各種の処理を実行する。RAM（Random Access Memory）13には、CPU11が各種の処理を実行する上において必要なデータやプログラムなどが適宜記憶される。入出力インタフェース15は、キーボードやマウスから構成される入力部16が接続され、入力部16に入力された信号をCPU11に出力する。また、入出力インタフェース15には、ディスプレイやスピーカなどから構成される出力部17も接続されている。

【0024】

さらに、入出力インタフェース15には、ハードディスクなどから構成される記憶部18、および、ネットワーク1を介して他の装置（例えば、端末3）とデータの授受を行う通信部19も接続されている。記憶部18には、ホームページに関するデータが記憶されており、他の装置から、管理しているホームページの提供の要請があった場合に提供するようになされている。ドライブ20は、磁気ディスク31、光ディスク32、光磁気ディスク33、半導体メモリ34などの記録媒体からデータを読み出したり、データを書き込んだりするときに用いられる。

【0025】

図3は、端末3の内部構成例を示す図である。端末3は、パーソナルコンピュータなどで構成することが可能であり、そのCPU41は、ROM42に記憶されているプログラムに従って各種の処理を実行する。RAM43には、CPU41が各種の処理を実行する上において必要なデータやプログラムなどが適宜記憶される。入出力インタフェース45は、キーボードやマウスから構成される入力部46が接続され、入力部46に入力された信号をCPU41に出力する。また、入出力インタフェース45には、ディスプレイやスピーカなどから構成される出力部47も接続されている。

【0026】

さらに、入出力インタフェース45には、ハードディスクなどから構成される記憶部48、インターネットなどのネットワークを介して他の装置（例えば、検

索サーバ4)とデータの授受を行う通信部49やドライブ50も接続されている。記憶部48には、WWWサーバ2から提供されるホームページを閲覧するために必要なブラウザなどのソフトウェアやデータが記憶されており、必要に応じ、読み出され、RAM43に展開される。

【0027】

図4は、検索サーバ4の内部構成例を示す図である。検索サーバ4は、パーソナルコンピュータなどで構成することが可能であり、そのCPU71は、ROM72に記憶されているプログラムに従って各種の処理を実行する。RAM73には、CPU71が各種の処理を実行する上において必要なデータやプログラムなどが適宜記憶される。入出力インタフェース75は、キーボードやマウスから構成される入力部76が接続され、入力部76に入力された信号をCPU71に出力する。また、入出力インタフェース75には、ディスプレイやスピーカなどから構成される出力部77も接続されている。

【0028】

さらに、入出力インタフェース75には、ハードディスクなどから構成される記憶部78、インターネットなどのネットワークを介して他の装置(例えば、端末3)とデータの授受を行う通信部79やドライブ80も接続されている。記憶部78には、WWWサーバ2により提供されるホームページを検索するためのデータが記憶されている。

【0029】

図5は、検索サーバ4の機能ブロック図である。検索サーバ4は、データを記憶する記憶機能と、その記憶されるデータを作成したり、記憶されているデータを用いた処理を実行する処理機能とから構成されている。検索サーバ4は、記憶機能として、データを収集するホームページ(サイト)のリストを記憶する収集サイトリスト記憶部101、収集サイトリスト記憶部101に記憶されているリストに基づき収集されたサイトのページのデータを記憶する保存ページ記憶部102、および、保存ページ記憶部102に記憶されたページデータが処理された結果を記憶するページデータ記憶部103を備えている。

【0030】

検索サーバ 4 は、処理機能として、保存ページ記憶部 102 に記憶されているページデータを処理するサイトページ処理部 111 と、サイトページ処理部 111 により処理された結果としてのデータを用いて所定の処理を実行し、関連ページに関するデータの生成などを行う関連ページデータ処理部 112 を備えている。

【0031】

サイトページ処理部 111 により処理されたデータは、ページデータ記憶部 103 のサイトページデータ記憶部 104 に記憶され、関連ページデータ処理部 112 により処理されたデータは、ページデータ記憶部 103 の関連ページデータ記憶部 105 に記憶される。

【0032】

サイトページ処理部 111 およびサイトページデータ記憶部 104 の詳細について、図 6 を参照して説明する。サイトページ処理部 111 は、ページ取得保存部 141 を備える。ページ取得保存部 141 は、収集サイトリスト記憶部 101 に記憶されているリストに記載されているサイトと接続する処理を実行し、各々のサイトに記憶されているホームページの全てのページのデータをダウンロードし、そのダウンロードしたデータを保存ページ記憶部 102 に記憶（保存）させる。

【0033】

保存ページ記憶部 102 に記憶されたページは、ページ ID 割り当て部 142 により、各ページが一意に区別がつくような ID が割り当てられ、その割り当てられた ID に関するデータが、サイトページデータ記憶部 104 のページ ID 記憶部 161 に記憶される。

【0034】

保存ページ記憶部 102 に記憶されたページは、単語抽出部 143 にも読み出される。単語抽出部 143 は、読み出したページ内から、そのページに含まれる単語を抽出する。単語抽出部 143 により抽出された単語のデータは、単語 ID 割り当て部 144 に供給される。単語 ID 割り当て部 144 は、供給された単語に対して、その単語が他の単語と区別がつくような ID を割り振る。その割り振

られたIDと、そのIDに対応する単語のデータは、サイトページデータ記憶部104の単語ID記憶部162に記憶される。

【0035】

単語割り当て部144からのデータは、基本ページモデル生成部145にも提供される。基本ページモデル生成部145は、抽出された単語が、そのページ内で、どのぐらいの頻度で用いられているかなどのデータを作成する。基本ページモデル生成部145により作成されたデータは、サイトページデータ記憶部104の基本ページモデル記憶部163に記憶される。

【0036】

保存ページ記憶部102に記憶されているページはサイトページ処理部111のリンク判定部146にも読み出される。リンク判定部146は、各ページの親子関係を判定する。各ページの親子関係とは、所定のページにおいて、そのページを親ページと称したとき、その親ページがリンクを張っている先のページを子ページと称したときの関係である。リンク判定部146により判定されたページ間の親子関係に関する情報は、サイトページデータ記憶部104のリンク情報記憶部164に出力され、記憶される。

【0037】

次に、関連ページ処理部112と関連ページデータ記憶部105の詳細な構成について、図7を参照して説明する。関連ページ処理部112は、必要に応じ、サイトページ記憶部104に記憶されているデータを用いて処理を実行する。まず、関連ページ処理部112のリンク関係情報生成部181は、サイトページデータ記憶部104に記憶されているデータを用いて同じ親ページを持つ子ページの情報を抽出する。

【0038】

図8を参照して説明するに、1つ所定の親ページからリンクが張られている子ページが複数存在している場合、その子ページの情報が抽出される。そして抽出された子ページ同士の情報、すなわち、兄弟(Sibling)と位置づけられるページ同士の情報が生成される。兄弟と位置づけられるページ同士の情報が、リンク関係情報生成部181において生成され、関連ページデータ記憶部105のリンク

関係情報記憶部 191 に記憶される。

【0039】

関連ページ処理部 112 の SDF データ生成部 182 は、SDF データを生成する。SDF とは、Sibling Document Frequency の略である。SDF データ生成部 182 により生成される SDF データとは、詳細は後述するが、各々のページに含まれる（各ページで出現する）単語に対して、その単語が現れる兄弟ページのリンクの重みを総和したデータである。

【0040】

SDF データ生成部 182 により生成された SDF データは、関連ページデータ記憶部 105 の SDF データ記憶部 192 に記憶される。関連ページ処理部 112 のページモデル拡張部 183 は、SDF データ記憶部 192 に記憶されているデータに対して重み付けを行い、その重み付けを行ったデータを、関連ページデータ記憶部 105 のページモデル拡張データ記憶部 193 に提供し、記憶させる。

【0041】

関連ページ処理部 112 の関連度算出部 184 は、ページ毎の関連度を算出し、その結果を、関連ページデータ記憶部 105 の関連度データ記憶部 194 に記憶させる。関連度算出部 184 が行う関連度の算出は、例えば、VSM (Vector Space Model の略、あるいは、ベクトル空間法と称される) の cosine 類似度に基づいて行われる。

【0042】

関連ページ一覧生成部 185 は、ユーザからの指示があった場合に、ページデータ記憶部 103 に記憶されているデータに基づいて、関連するページについての一覧表を作成し、そのデータを提供するという処理を実行する。

【0043】

このようなデータを生成し、記憶する検索サーバ 4 と端末 3 との間で行われる処理について、図 9 のフローチャートを参照して説明する。ステップ S11 において、端末 3 は、ネットワーク 1 を介して検索サーバ 4 に接続される。ここでの接続（アクセス）は、始めて端末 3 が検索サーバ 4 と接続されたとき、又は、端

末 3 側で後述する設定が行われていないときであるとする。換言すれば、ユーザが検索を行うために、後述する関連ページ検索ボタン 231（図 10B）を操作したときの接続とは異なる接続であるとする。

【0044】

検索サーバ 4 は、端末 3 からのアクセスを受け付けると、ステップ S 21 において、導入画面の送付を行う。導入画面とは、端末 3 のユーザが、検索サーバ 4 による検索を行う際に操作するボタンなどを、端末 3 のブラウザ上に設定するための画面であり、例えば、図 10A に示したような画面である。

【0045】

端末 3 の記憶部 48（図 3）には、ネットワーク 1 を介してデータの授受を行う際に用いられるブラウザに関するプログラムが記憶されており、必要に応じ、起動され、CPU 41 が処理を実行する際に用いられる。ブラウザが起動され、検索サーバ 4 からの導入画面のデータが受信され、起動されているブラウザにより処理されると、図 10A に示したような画面が、出力部 47 としてのディスプレイ 211 上に表示される（ステップ S 12）。

【0046】

ディスプレイ 211 には、ブラウザが起動されることにより表示される部分の下側に、画像表示部 221 が設けられており、その画像表示部 221 に、検索サーバ 4 からの導入画面が表示される。導入画面としては、例えば、“このボタンをドラッグアンドドロップすると、関連ページ検索エンジンがブラウザ上に設定されます”といったメッセージと共に、ボタンが表示されている画面である。ユーザは、このメッセージに従って、ボタンを、例えば、ブラウザの上部の所定の欄（通常、リンクツールバーという欄）にドラッグアンドドロップを行う。

【0047】

このようなドラッグアンドドロップが、ステップ S 13 において行われると、そのドラッグアンドドロップの処理に対応する設定が、ステップ S 14 において、行われる。すなわち、例えば、図 10B に示したように、ドラッグアンドドロップされたボタンに対応する関連ページ検索ボタン 231 が、ブラウザの所定の部分に表示され、その関連ページ検索ボタン 231 に関連付けられて、検索サー

バ4のアドレスが記憶されるなどの設定である。

【0048】

このような設定が行われることにより、図10Bに示したように、ブラウザ上の所定の部分に関連ページ検索ボタン231が表示されるようになると、ユーザは、検索サーバ4による検索を利用することが可能な状態とされる。

【0049】

このような導入画面を用い、関連ページ検索ボタン231がブラウザ上に設定されるようにしても良いし、所定のページ内に、バナーとして関連ページ検索ボタン231が設けられているようにしても良い。また、ユーザが検索サーバ4にアクセスし、所定のページのURL (Uniform Resource Locationの略) を入力することも可能である。いずれにしても、ユーザが検索を所望したときに、ボタンのクリックなど簡便な操作で、検索サーバ4にアクセスでき、その検索サーバ4による検索の結果を授受できるように設定されていればよい。

【0050】

ここでは、図10Bに示したように、ブラウザ上に、関連ページ検索ボタン231が設定されているとして説明する。ユーザが、所定のページ、例えば、WWWサーバ2-1 (図1) により提供されているホームページの所定のページを閲覧している状態で、関連ページ検索ボタン231を操作すると、検索サーバ4に、関連ページ検索ボタン231が操作されたという情報、すなわち、検索が指示されたという情報が送信される。その結果、検索サーバ4においては、図11に示すようなフローチャートの処理が開始される。

【0051】

ステップS41において、所定のホームページ (サイト) のページのデータ (以下、単にページと記述した場合も、ページのデータという意味を示すとする) が取得され、保存される。取得されるホームページのページは、収集サイトリスト記憶部101に記憶されているリストに基づくものである。また、ユーザからの要求により送信された所定のURLが収集サイトリスト記憶部101に記録されていない場合には、そのURLを追加し記録する。収集サイトリスト記憶部101に記憶されているリストの一例を図12に示す。図12に示したように、収

集サイトリスト記憶部 1 0 1 に記憶されているリストには、“収集開始 URL”、“含むディレクトリ”、“排他ディレクトリ”、“含むドメイン”、および、“排他ドメイン”といった情報が含まれる。

【 0 0 5 2 】

このようなリストに基づいて、ページが取得される。取得されたページは、保存ページ記憶部 1 0 2 に保存され記憶される。また、保存ページ記憶部 1 0 2 には、図 1 3 に示すようなリスト形式で、取得されたページのサイト単位での情報が管理されている。図 1 3 に示したように、リストには、“サイト ID”、“サイト名”、および、“総ページ数”といった情報が含まれる。

【 0 0 5 3 】

サイト ID は、そのサイトに割り当てられた ID であり、ページ取得保存部 1 4 1 がページ（サイト）の情報を取得した時点で、割り当てるようにしても良いし、収集サイトリスト記憶部 1 0 1 で記憶されている図 1 2 に示したようなリストで、ID も関連付けて記憶するようにし、その ID が、記憶されるようにしても良い。

【 0 0 5 4 】

このようにして、保存ページ記憶部 1 0 2 に取得されたページが保存され、所定のサイトの情報が記憶されると、ステップ S 4 2 において、ページ ID 割り当て部 1 4 2 により、取得されたページ毎に ID が割り当てられる。ページ ID 割り当て部 1 4 2 は、保存ページ記憶部 1 0 2 に記憶されているページを読み出し、そのページに ID を割り当てる。

【 0 0 5 5 】

この際、読み出されるページと、割り当てられた ID から、図 1 4 に示したようなリストが作成され、ページ ID 記憶部 1 6 1 に記憶される。図 1 4 に示したページ ID 記憶部 1 6 1 に記憶されるリストには、“ページ ID”、“サイト ID”、“ページ URL”、“タイトル”、“サマリー”、“ページ保存場所”、および、“最終更新日”といった情報が含まれる。

【 0 0 5 6 】

これらの情報のうち、“ページ ID”は、ページ割り当て部 1 4 2 により割り

当てられ、その他の情報は、保存ページ記憶部 102 に記憶され、読み出されたページのデータから抽出される。

【0057】

ステップ S43 において、ページ内に含まれる単語が、単語抽出部 143 により抽出される。この単語の抽出は、単語抽出部 143 が、保存ページ記憶部 102 から保存されているページのうちの 1 ページを読み出し、そのページに含まれている単語が抽出されることにより行われる。抽出される単語としては、名詞に分類される単語である。形容詞や動詞などに分類される単語や、英語なども抽出されるようにしても良い。単語抽出部 143 により抽出される単語は、後の処理において必要となる（検索サーバ 4 が最終的に検索結果としてユーザに提供する結果が良好になるために必要となる）品詞の単語が抽出できれば良い。

【0058】

抽出された単語は、単語 ID 割り当て部 144 に供給される。単語 ID 割り当て部 144 に供給されるのは、抽出された単語だけでなく、その単語の出現回数、ページ ID、タグ付きの各単語、そのタグ付きの単語の出現回数なども供給される。これらの情報は、必要に応じ、単語抽出部 143 が、ページ ID 記憶部 161 や保存ページ記憶部 102 から読み出し、供給する。

【0059】

単語 ID 割り当て部 144 は、供給された単語に対して ID を割り当てる。ID が割り当てられた単語は、ID と関連付けられて、単語 ID 記憶部 162 に記憶される。単語 ID 記憶部 162 には、例えば、図 15 に示したようなリストが記憶されている。

【0060】

図 15 に示したように、単語 ID 記憶部 162 には、“単語 ID” と “単語” が関連付けられて記憶されている。なお、同一の単語が抽出された場合には、同一の ID が割り当てられる。そのために、単語抽出部 143 は、抽出された単語が、既に単語 ID 記憶部 162 に記憶されている単語であるか否かを判断し、既に記憶されている単語である場合には、新たに ID を割り振ることが無いように制御されている。

【0061】

また、単語ID割り当て部144は、図16に示したようなリストを作成し、単語ID記憶部162に記憶させる。図16に示したリストは、“単語ID”、“サイトID”、“そのサイト内で当該単語を含むページ数”、および、“そのサイト内で当該単語を含むページ”といった情報を含むものである。図16に示したリストは、所定の1つのサイトに注目したときに、そのサイトに含まれる所定の単語との関連を示すものである。

【0062】

単語IDと、その単語IDが割り当てられた単語は、単語ID記憶部162に供給されると共に、その一部のデータは、基本ページモデル生成部145にも供給される。基本ページモデル生成部145は、ステップS45において、基本ページモデルを生成する。基本ページモデルとは、図17に示したようなデータであり、基本ページモデル記憶部163に記憶されるリスト形式のデータである。このようなデータを作成するために、基本ページモデル生成部145は、単語ID割り当て部144から、ページIDと、それぞれの単語IDとその出現回数に関する情報が供給される。

【0063】

図17に示すように、基本ページモデル記憶部163に記憶されるリストは、“ページID”、“出現単語”、“Title”、“Keywords”、および、“description”といった情報が含まれる。このリストは、1つのページに対して、1つの単語が、何回出現しているか（用いられているか）を示す情報であり、また、タイトル（Title）などの種類毎に分類された情報も含む情報である。このような種類毎に分類された情報は、最終的に関連するページを決定する際に、単語の用いられている部分毎（種類毎）に重要度が異なることが考えられ、そのような重要度の違いにより重み付けを行うようにした場合のことを考慮したときに用いられる。

【0064】

ステップS46において、リンク判定部146は、図8を参照して説明したように、親ページと、そのページがリンクしている子ページを判断し、その判断結

果をリンク情報記憶部 164 に記憶させる。リンク情報記憶部 164 に記憶されている情報は、例えば、図 18 に示したような情報である。

【0065】

図 18 に示したように、リンク情報記憶部 164 に記憶されているリスト形式の情報は、“ページ ID”、“リンク先のページ ID”、“リンクの重み”、および、“アンカー窓内単語”といった情報が含まれている。“ページ ID”と、“リンク先のページ ID”、すなわち、親としてのページと子としてのページが関連付けられていることがわかる。このような情報を作成するために、リンク判定部 146 は、必要に応じ、保存ページ記憶部 102、ページ ID 記憶部 161、および、基本ページモデル記憶部 163 からデータを読み出す。

【0066】

“リンクの重み”は、以下のようにして算出される。なお、重み付けを算出する際、アンカー窓内に含まれる単語が、リンク先のページ（この場合、子ページ）に含まれるほど、ページ間の関連性が高いと考え、重みを増加するようにする。また、リンク元のページ（すなわち、親ページ）が多くのリンクを持つほど、1つのリンクに対する重要性は低いと考えられるため、そのようなページからリンクが張られている子ページとのリンクの重みは小さくなるようにする。

【0067】

親ページ p から子ページ q へのリンクの重み $W_c(p, q)$ は、次式 (1) に基づいて算出される。

$$W_c(p, q) = 1 + N_{pq}(T_{anc}) \times 1/k \quad \dots (1)$$

式 (1) において、 $p, q \in P$ (P はページ集合) である。また、 $N_{pq}(T_{anc})$ は、親ページ p 内のアンカー窓内の単語の集合を集合 (T_{anc}) とし、その集合 (T_{anc}) の子ページ q 内の出現数を表している。なお、 $T_{anc} \in T_{all}$ であり、 T_{all} は、全単語の集合とする。

【0068】

k は、親ページ p が有しているリンクの数であり、ページ p からページ q へのリンクを含むため、 k は、常に 1 以上の数に設定される。なお、式 (1) において、右辺の第 1 項で 1 だけ加算しているのは、算出される重み $W_c(p, q)$ が 1

未満にならないようにするためである。

【0069】

このようにして重み $W_c(p, q)$ が算出されるようにしても良いし、アンカー窓内の出現単語にアンカーを中心とした距離に応じた重み付けを行って $W_c(p, q)$ が算出されるようにしても良い。アンカーを中心とした距離に応じた重み付けを行って $W_c(p, q)$ を算出するようにした場合、式 (1) における $N_{pq}(T_{anc})$ は、次式 (2) に基づいて算出される。

$$N(p, q)(T_{anc}) = H(Dis(t_1)) \times T_c(t_1) + H(Dis(t_2)) \times T_c(t_2) + \dots + H(Dis(t_k)) \times T_c(t_k) \quad \dots (2)$$

【0070】

式 (2) において、 $tk \in T_{anc}$ であり、 $Dis(tk)$ は、アンカータグから単語 tk が出現するまでの距離を示し、 $0 \leq Dis(tk) \leq D_{max}$ の値を取る。 D_{max} は、アンカー窓の片方の最大幅である。また、 $H(Dis(tk))$ は、 $Dis(tk)$ に対する重みを表し、 $0 < H(Dis(tk)) \leq 1$ の範囲内の値であり、 $H(0) = 1$ である。 $T_c(tk)$ は、単語 tk の子ページ q 内の出現数を表す。

【0071】

このように、アンカー窓からの距離を考慮した重み付けを行うようにしても良い。また、アンカー窓内の単語のアンカー窓内出現数や、リンク先ページ（子ページ）での出現数にタグの種類に応じた重み付け（重要度）を考慮した重み付けを行うことも可能である。また、これらの重み付けを行わず、単に $W_c(p, q) = 1$ としてもよい。

【0072】

このようにして、図 18 に示したリンク情報記憶部 164 に記憶されるリスト内の“リンクの重み”は算出される。図 11 のフローチャートの説明に戻り、ステップ S47 において、リンク関係情報の生成が、リンク関係情報生成部 181（図 7）により行われる。リンク関係情報生成部 181 により作成された情報は、リンク関係情報記憶部 191（図 7）に、図 19 に示したようなリスト形式で記憶される。リンク関係情報生成部 181 は、図 19 に示したような情報を作成するための情報を、リンク情報記憶部 164 から取得する。

【0073】

図19に示したように、リンク関係情報記憶部191には、“ページID”、“SiblingページID”、および、“リンクの重み”が、それぞれ関連付けられて記憶されている。ここで、Siblingページとは、共通の親ページを有する子ページのことであり、図8を用いて説明したように、兄弟関係にあるページのことを示す。

【0074】

リンク関係情報生成部181は、各々のページIDに対して、Siblingの関係にあるページIDを抽出するといった処理を行うとともに、Siblingページ間のリンクの重みも算出する。そのSiblingページ間のリンクの重みの算出は、以下のようにして行われる。すなわち、Siblingページ間のリンクの重み $W_s(r,s)$ は、次式(3)に基づいて算出される。

【0075】

$$W_s(r,s) = W_c(t,r) \times W_c(t,s) \quad \cdots (3)$$

式(3)において、 r,s,t は、 P をページ集合とした場合、 $r,s,t \in P$ を満たす値であり、 $W_s(r,s)$ は、 $1 \leq W_s(r,s)$ を満たす値である。

【0076】

式(3)において、 $W_s(r,s)$ は、所定のページ r と、そのページ r とSiblingの関係にあるSiblingページ s 間のリンクの重みであり、 $W_c(t,r)$ は、所定のページ t と、そのページ t と親子関係にある子ページ r 間のリンクの重みであり、 $W_c(t,s)$ は、所定のページ t と、そのページ t と親子関係にある子ページ s 間のリンクの重みである。

【0077】

図20を参照して、式(3)について説明するに、この重みの算出は、所定のページ r と、そのページ r とSibling(兄弟)の関係にあるページ s とのリンクの重み $W_s(r,s)$ は、そのSibling関係内に存在するページ間のリンクの重み、この場合、ページ r とページ s とに共通に親子の関係にある親ページ t とのリンクの重みである、重み $W_c(t,r)$ と重み $W_c(t,s)$ とを乗算することにより求められる。

【0078】

このようにして、Siblingページ間のリンクの重みが算出され、その算出結果が、図19に示したようなリスト形式のデータに書き込まれる。

【0079】

図11のフローチャートの説明に戻り、ステップS48において、SDFデータの生成が、SDFデータ生成部182（図7）により行われる。SDFデータ生成部182は、必要に応じ、リンク関係情報記憶部191と基本ページモデル記憶部163からデータを読み出し、その読み出したデータを用いて、図21に示すようなリスト形式のデータを作成し、SDFデータ記憶部192に記憶させる。

【0080】

図21に示したSDFデータ記憶部192に記憶されるデータは、“ページID”と、“ページIDに含まれる単語IDと、その単語IDを含むSiblingページのリンクの重みの総和”といった情報を含む。このデータは、各々のページにおいて、そのページ内で出現する単語に対して、その単語が現れるSiblingページのリンクの重みを総和したデータであり、リンク判定部146が、前述のとおり $W_c(p, q) = 1$ とリンクの重みを生成した場合には、単にその単語が現れるSiblingページの総数となる。

【0081】

ステップS49において、ページモデル拡張部183（図7）は、ページモデル拡張処理を実行する。ページモデル拡張処理とは、図22に示すようなリスト形式のデータを作成し、ページモデル拡張データ記憶部193に記憶させる処理である。ページモデル拡張部183は、図22に示すようなデータを作成するために、基本ページモデル記憶部163、リンク情報記憶部164、リンク関係情報記憶部191、および、SDFデータ記憶部192に、それぞれ記憶されているデータを必要に応じて読み出す。

【0082】

図22に示したページモデル拡張データ記憶部193に記憶されているデータは、“ページID”と“ベクトル”といった情報を含む。“ベクトル”内の重み

は、ISDF (Inverse Sibling Document Frequency) に基づき、以下のよう
して求められる。

【0083】

$$P_i = (\{T_{i1} \times W_{i1}\}, \{T_{i2} \times W_{i2}\}, \dots, \{T_{ij} \times W_{ij}\}, \dots \dots (4)$$

式(4)において、 i はページであり、 $i \in P$ 、 j は単語であり、 $j \in T_{all}$ である。
 P_i は、ページ i の T_{all} 次元のベクトルを示す。 T_{ij} は、ページ i において単語 j が
出現しているか否かを示す値であり、出現している場合1が、出現していない場
合0が、それぞれ設定される。

【0084】

W_{ij} は、ページ i における単語 j の重みであり次式(5)に基づいて算出される
。また、 W_{ij} は、 $0 \leq W_{ij}$ を満たす値となり、 $\sum (T_i \times W_{ij})^2 = 1$ (T_i と W_{ij}
を乗算した値を2乗した値の総和が1)になるよう正規化される。

【0085】

$$W_{ij} = (1 + \log(TF_{ij})) \times (1 + \log(1 / (1 + SDF_{ij}))) \dots (5)$$

式(5)において、 TF_{ij} は、単語 j のページ i における出現回数を示し、 $0 \leq TF_{ij}$
の値を取る。 SDF_{ij} は、ページ i のSiblingページのうち、単語 j を含むページ
のリンクの重みの総和を示す。

【0086】

このような式(4)と式(5)を用いてベクトル内の重みを算出するようにし
ても良いが、さらに、 SDF_{ij} の効果を高めるため、式(5)を式(6)に置き換
えても良い。

$$W_{ij} = (1 + \log(TF_{ij})) \times (1 + \log(1 + ASDFi / (1 + SDF_{ij}))) \dots (6)$$

【0087】

式(6)において、 $ASDF_i$ は、ページ i と全Siblingページ間のリンクの重みの
総和を示す。

【0088】

さらに、 TTF_{ij} 、 ATF_{ij} を加え、式(5)を基に次式(7)あるいは、式(6)
を基に次式(8)に基づいて、重みを算出するようにしても良い。

$$W_{ij} = (1 + \log(TF_{ij} + TTF_{ij} + ATF_{ij})) \times (1 + \log(1 / (1 + SDF_{ij}))) \dots (7)$$

$$W_{ij} = (1 + \log(TF_{ij} + TTF_{ij} + ATF_{ij})) \times (1 + \log(1 + ASDF_i / (1 + SDF_{ij}))) \dots (8)$$

式(7)、(8)において、 TTF_{ij} は、タグ付単語jがページiにおいて出現する
 可否を示し、出現しない場合0が、出現する場合1が、それぞれ設定される。
 あるいは出現回数(0以上)が設定されるようにしても良い。タグの種類に応じて
 それぞれ重みを付けるようにしても良い。

【0089】

また、 ATF_{ij} は、ページiのリンク元ページ（この場合、親ページ）におけるア
 ンカー窓内において単語jが出現する可否を示し、出現しない場合0が、出現す
 る場合1が、それぞれ設定される。あるいは出現回数(0以上)が設定されるよう
 にしても良い。タグ付単語と同様に、重みを付けてもよい。さらに、アンカーか
 らの距離に応じて重みをつけるようにしても良い。

【0090】

このような式に基づいて、図22に示したデータ内の“ベクトル”内の各々の
 単語に対する“重み”のデータが算出される。図11のフローチャートの説明に
 戻り、ステップS50において、関連度算出部184において、ページ間の関連
 度が算出される。関連度算出部184は、必要に応じ、ページモデル拡張データ
 記憶部193に記憶されているデータを読み出し、図23に示すようなリスト形
 式のデータを作成し、関連度データ記憶部194に記憶させる。

【0091】

図23に示した関連度データ記憶部194に記憶されるデータは、“ページI
 D”、“対象ページID”、“関連度”、および、“高関連度単語”といった情
 報を含む。このうち、関連度は、以下のようにして算出される。関連度は、関連
 ページ検索に適した形に特徴抽出されたページ間の特徴が共通している部分が多
 いほど、関連度が高くなるという考えに基づき算出される。例えば、共通特徴数
 ／総特徴数（積／和）、VSMのcosine類似度などを用いて算出することができ
 る。

【0092】

具体的には、次式(9)に基づいて算出される。式(9)は、VSMのcosine類似度によるものである。

$$R(i, j) = P_i \cdot P_j / (|P_i| |P_j|) \cdots (9)$$

式(9)において、 P_i 、 P_j はそれぞれページ*i*、ページ*j*のベクトル表現であり、式(4)により算出(表現)される値である。また、 $i, j \in P$ である。 $R(i, j)$ は、ページ*i*に対するページ*j*の関連度であり、図23では、ページ*i*が“ページID”，ページ*j*が“対象ページID”となる。

【0093】

このようにして算出された関連度は、図23に示すようなリスト形式のデータ内のデータとして、関連度データ記憶部194に記憶される。次に、ステップS51以降の処理が行われるわけだが、ステップS51以降の処理は、このようにして各記憶部に記憶されたデータ、特に、関連度データ記憶部194に記憶されているデータが用いられて行われる。

【0094】

そこで、ここまでの処理、すなわち、ステップS41乃至S50までの処理は、ユーザの要求があった際に、リアルタイムに実行されるようにしても良いし、ユーザの要求に関わらず、事前に実行されるようにしても良い。

【0095】

ユーザの要求に関わらず、ステップS41乃至S50の処理が行われる場合、所定のサイトから定期的にデータを取得するようにし、各記憶部に記憶されているデータが更新されるようにすれば良い。このように、予めデータを作成しておけば、ユーザからの要求があった際、ユーザからの要求があってからリアルタイムに処理を実行するよりも、その要求に即座に対応することが可能となる。

【0096】

また、上記のように予めデータを作成した場合、ユーザから要求がある際に送信されるURLが予め作成したデータに存在しないときには、ステップS41乃至S50をそのURLの示すページ、あるいはそのページのサイトについて行うことが可能である。

【0097】

ステップS51において、関連ページ一覧生成部185は、ユーザが関連ページの提供を指示してきたページに対応する関連ページの一覧を作成する。その作成は、以下のようにして行われる。

【0098】

まず、関連ページ一覧生成部185は、ページID記憶部161から、ユーザが関連ページ検索ボタン231を操作した際に閲覧されていたページ（関連ページの検索が指示されたページ）のURLに対応するページIDを読み出す。その読み出されたページIDをKey1とするデータが、関連度データ記憶部194（図23）から読み出される。その際、関連度の値が高い順にソートされ、その関連度に該当する対象ページID（Key2となるページID）が読み出される。

【0099】

そして、関連ページ一覧生成部185は、該当したページIDをページID記憶部161に照合し、URLなど、そのページに関する情報を取り出し、一覧データを生成する。

【0100】

一覧データを生成する際、ここまでの処理により得られたデータで終了しても良いが、さらに、以下のような機能を付け加えても良い。ユーザには、関連度が高い順にページに関する情報が表示されるように、一覧表が作成されるわけだが、例えば、同一の関連度を有するページが複数存在する場合は考えられ、そのようなとき、どのページを上位に表示するかが問題となる。また、関連度とは関係しないページの重要度を加味して、最終的にユーザへ関連ページを表示することも考えられる。

【0101】

そこで、関連度算出部184が算出した関連度に対して、ページのランク付けを行い、そのデータを最終的な関連度の値に付加するようにする。例えば、ページのランク付けとしては、検索サーバ4自体が、ランク付けの機能を有するようにしても良いし、他のサーバで提供しているランク付けの情報を引用するように

しても良い。

【0102】

ランク付けのデータを加味した関連度の算出は、具体的には、パラメータによる調整が考えられる。

$$R'(i, j) = p R(i, j) + (1-p) G(j) \quad \cdots (10)$$

式(10)において、 $R'(i, j)$ は、ページ*i*に対するページ*j*のランク付関連度であり、 $R(i, j)$ は、ページ*i*に対するページ*j*の関連度であり、式(9)により算出される値である。また、 $G(j)$ は、ページ*j*のランクであり、 p は、 $0 \leq p \leq 1$ の値を有するパラメータである。この式(10)で算出されたランク付関連度をすでに述べた図23に示すようなリスト形式のデータ内のデータとして、関連度データ記憶部194に記憶してもよい。

【0103】

また、上述した実施の形態でステップS49において、ページモデル拡張部183が行う処理の前または後の処理として、リンク先のページを考慮したページモデルを作成するようにしても良い。具体的には、所定のページの基本ページモデルに、リンク先の基本ページモデルの総和を付加する。このようにした場合、上述したリンク判定部146で算出されるリンク間の重みを付加するようにしてもよい。最下層（葉）のページまで計算する、あるいは、N回のリンク先まで考慮という形にする。

【0104】

ISDFによるページモデル拡張部183が行う処理の前に、この機能を実現した場合、所定のページのページモデルに存在する単語種が増えるため、ISDFの結果が影響を受けることになるため、このことを考慮して、前または後の、どちらに処理を実行するかを決定した方が良い。

【0105】

さらに、上述した実施の形態において、各処理を行う上で、単語の関連性ということ considering して処理を行うようにしても良い。例えば、“旅行”と“海外”といった単語を関連付けた辞書（関連辞書）を設け、その関連辞書を参照して処理が行われるようにする。このような関連辞書を設けない場合は、ページ内に出現

した単語のみで関連度が決定されるが、関連辞書を設けるようにした場合は、例えば、基本ページモデル生成部145やSDFデータ生成部182、あるいは関連度算出部184などが処理を実行する前の処理として、関連辞書が参照され、その結果が用いられて関連度が算出されるようにしても良い。関連辞書としては、共起情報やKeyGraph手法により作成されるか、ODP（Open Directory Projectの略）のカテゴリー情報などが利用されるようにしても良い。

【0106】

図11のフローチャートの説明に戻り、このようにして生成された一覧データは、ステップS52において、ネットワーク1を介して端末3に送信される。端末3側において、一覧データが処理されることにより、ユーザに関連ページの一覧表が提供される。この関連ページの一覧表は、端末3のディスプレイ211上では、既に開かれているウインドウ（関連ページ検索ボタン231が操作されたウインドウ）とは異なるウインドウとして表示されるようにしても良いし、既に開かれているウインドウに表示されるようにしても良い。

【0107】

ここで、このような検索サーバ4による検索の結果として、ユーザに提供される関連ページについて説明する。例えば、従来の手法により所定のページの関連ページを検索した場合、その検索される関連ページは、類似しているページが上位に表示されるようになっていた。例えば、所定のミュージシャンのサイト内のプロフィールのページを閲覧しているときに、そのページに関連するページを検索した場合、そのミュージシャンの他のサイト内のプロフィールのページが検索結果としてユーザに提供されるといったことが行われていた。

【0108】

しかしながら、この例の場合、同一のミュージシャンの同一のプロフィールを、別のサイトで閲覧してもユーザにとって、新たに得られる情報は無いといえる。換言すれば、ユーザは、同一のミュージシャンのプロフィールを何度も閲覧したいわけではなく、プロフィールに関連する情報、例えば過去に参加したイベントに関する情報や、プロフィールに記載されたストーリーに関する情報、ミュージシャンが好む事柄に関する情報などを所望しているために、関連ページの

検索を実行したと考えられる。すなわち、ユーザは、検索を実行する際、重複した情報である類似するページを参照したいわけではなく、何らかの関わりのあるページを参照したいと考えられる。このような、類似しているわけではないが、関連しているページを提供することが、上述した検索サーバ4による検索においては実現することが可能である。

【0109】

上述した検索サーバ4の処理を図24を参照して説明する。図24に示すように、親ページには、リンクが張られている子ページとして子ページ1乃至3が存在するとする。そして、子ページ1に含まれる単語（ステップS43の処理で抽出される単語）が、“a, b, c, . . .”であり、子ページ2に含まれる単語が、“a, c, d, . . .”であり、子ページ3に含まれる単語が、“a, x, . . .”であるとする。

【0110】

このような状況では、子ページ1乃至3には、共通に、単語aが含まれている。例えば、所定の会社が運営するサイトの所定の製品Aのホームページ内で、使い方の提案などが掲載されているページがあるとする。そのページ内には、製品Aの名称を示す単語aが、高い確率で含まれている可能性がある。そのような場合には、単語aは、各ページの特徴を示す単語として（他のページとの差異を表す単語として）は、ふさわしくないと考えられる。

【0111】

よって、単語aなど、複数のページに共通に含まれる単語などは、それらのページの特徴を表す単語として取り扱われないようにする。換言すれば、ページ間の関連度を判断するためのページの特徴抽出としては、単語aなど、複数のページに共通に含まれる単語などは、他の単語と比較して重要度が低く設定される（他の単語の方が、重みが重く設定される）ようにする。

【0112】

その重みの設定は、上述したように、本実施の形態においては、ISDF（Inverse Sibling Document Frequency）に基づいて行っている。このISDFに基づく重み付けは、上述したように、ステップS49の処理として、ページモデル

拡張部 183 (図 7) が行っている。

【0113】

ここで、従来の重み付けの手法として、TF-IDF (Term Frequency-Inverse Document Frequency) がある。重み付けに TF を用いるのは、文書中 (所定のページ中) で繰り返し用いられる単語は、そのページ内において重要な概念であると考えられるためである。しかしながら、ページ内に多く用いられている単語の中には、そのページを特定する性質を持たない共通あるいは汎用の単語も多く、索引語として適していないことが多い。そこで、語がどのくらい特定性を持つかを IDF によって重み付けに反映させるという手法である。

【0114】

IDF により、所定のデータセットの多くの文書に出現する単語の重みを小さくする効果が得られる。そのため、所定のデータセット内のページの特徴をより明確に出すことが可能となる。

【0115】

この TF-IDF の IDF に対して、本実施の形態においては、ISDF という手法を用いている。従って、本実施の形態においては、TF-ISDF という手法を用いて重み付けを行っていることになる。これは、TF-IDF の手法と異なり、所定の関係 (この場合、兄弟関係にあるページであり、後に詳細を示す ICDF では、共通親関係) の文書群を 1 つのデータセットとみなし、IDF を適用していると考えられる。

【0116】

すなわち、何を共通のデータセットとして見なすかが異なることになる。本実施の形態においては、兄弟関係にある文書 (ページ) を 1 つのデータセットと見なしている。この兄弟関係にあるページとは、リンク元のページが共通という関係にある。リンク元のページが共通という関係にあるということは、そのページ間において、何らかの関係がある、何らかの類似点 (共通点) があると考えられる。

【0117】

そのような類似点 (共通点) があるページ群を 1 データセットとみなし、重み

付けを行う（ISDFに基づく処理を行う）ことにより、類似したページの間の差分が、より明確になると考えられる。これにより、関連ページ検索に適した形で、各ページの特徴をより明確にすることになると考えられる。

【0118】

このようなことを換言すれば、どこまでを不要な特徴（雑音）として見なし、排除するかを適切に設定することにより、類似する文書に含まれる単語の重みを減じ、それらの文書（ページ）の他の特徴を浮き出させる。このように他の特徴を浮き出させることにより、類似度ではなく、関連度を求めるためのページの重み付け（特徴抽出）を行うことが可能となる。

【0119】

つまり、TF-IDFのIDFは、あるデータセット内のページに共通で用いられる単語を不要な特徴とみなし、各ページの特徴を明確にすることで、キーワードを入力し検索結果を出力する従来の検索エンジンに適したページの特徴抽出方法として用いられてきた。しかしながら、TF-ISDFのISDFは、類似点がある兄弟関係のページ群をデータセットとみなし、その中で共通に用いられる単語を不要な特徴とみなすことで、関連ページ検索に適した特徴抽出手法であるといえる。

【0120】

このような重み付けが行われた結果が用いられて、関連度が、例えば、VSMのcosine類似度などに基づいて算出される。この関連度の算出は、上述した実施の形態においては、関連度算出部184により行われる。VSMについて簡便に説明するに、VSMによる手法は、出現する単語の有無や出現数を特徴量とし、検索対象データや入力文書を全単語次元数のベクトルで表現するものである。VSMでは、データ間の類似度（共通する度合い）を算出するために、ベクトル間のcosineを用いることが多い。VSMによる手法は、記事と語彙の関係、記事同士の関係、単語同士の関係をモデル化するのに有効な手法とされている。

【0121】

本実施の形態において、上述したような重み付けを行い、関連度を算出し、その関連度を用いて、ユーザに対して関連ページの情報を提供するため、例えば、

所定のミュージシャンのサイト内のプロフィールのページを閲覧しているときに、そのページに関連するページを検索した場合、そのミュージシャンの他のサイト内の同一プロフィールのページが検索結果としてユーザに提供されるというのではなく、そのミュージシャンの過去に参加したイベントに関する情報や、プロフィールに記載されたストーリーに関する情報、ミュージシャンが好む事柄に関する情報などの情報がユーザに提供されることになる。

【0 1 2 2】

従って、本実施の形態によれば、ユーザが所望する関連ページをより高い精度で提供することが可能となる。

【0 1 2 3】

一方、本実施の形態における、ページの兄弟関係、あるいは詳細を後述する共通親関係を用いたページの特徴抽出手段は、ユーザのブラウジング履歴のなかの所定のページを用いたユーザモデル生成法に適用可能である。すなわち、ユーザモデルの生成法は、ユーザが過去に参照したページ群を解析することによって生成されることが多いが、そのページの特徴抽出手段として、本実施の形態にある兄弟あるいは共通親関係のページを考慮したページの特徴抽出手段が利用できる。さらに、キーワードや自然言語を入力とした検索エンジンへ適用し、兄弟関係、あるいは共通親関係を考慮したページモデルに基づく検索エンジンの実現も可能である。

【0 1 2 4】

上述した実施の形態においては、リンク判定部 1 4 6（図 6）は、親ページに注目して、その親ページがリンクを張っている他の子ページを判定するようにし、その結果を用いて後段の処理が行われるとしたが、子ページに注目して、その子ページにリンクを張っている他の親ページを判定するようにし、その結果を用いて後段の処理が行われるようにしても良い。

【0 1 2 5】

すなわち、図 2 5 を参照して説明するに、所定の子ページに注目した際、その子ページにリンクを張っている複数の親ページ（共通親のページ）が存在している場合が考えられ、それらの共通親 (Co-Parent) ページの関係を、リンク関係情

報生成部 181 に相当する部分が判定し、その判定結果が用いられて、後段の処理が行われるようにしても良い。

【0126】

そのような判定結果を用いるようにした場合について説明する。検索サーバ 4 の内部構成は、基本的に、図 5 乃至図 7 に示したような構成と同様に構成することが可能である。ただし、図 7 に示した部分に関する構成は、図 26 に示したような構成となる。図 7 に示した構成と、図 26 に示した構成とを比較するに、図 26 に示した構成は、図 7 の SDF データ生成部 182 と SDF データ記憶部 192 を、それぞれ CDF データ生成部 252 と CDF データ記憶部 262 に置き換えた構成とされ、他の部分は、同じ構成とされている。しかしながら、各部で処理されるデータが異なり、その異なる部分について、以下に説明する。

【0127】

図 26 に示した構成を含む検索サーバ 4 の動作は、図 27 に示したフローチャートの処理に従って行われる。ここで、図 27 に示したフローチャートを参照して、図 26 に示した構成を含む検索サーバ 4 の動作について説明する。ステップ S71 乃至 S76 の処理は、図 11 に示したフローチャートのステップ S41 乃至 S46 の処理と同様の処理であるので、その説明は省略する。

【0128】

ステップ S71 乃至 S76 における処理、すなわち、検索サーバ 4 内の構成のうち、図 6 に示した部分で行われる処理が行われることにより、図 6 に示した、保存ページ記憶部 102、ページ ID 記憶部 161、単語 ID 記憶部 162、基本ページモデル記憶部 163、および、リンク情報記憶部 164 にはそれぞれ、図 14 乃至図 18 に示したデータが記憶される。

【0129】

ステップ S77 において、リンク関係情報が、リンク関係情報生成部 251 により生成されるわけだが、その生成され、リンク関係情報記憶部 261 に記憶されるデータは、図 28 に示したようなデータである。図 28 に示したように、リンク関係情報記憶部 261 には、“ページ ID”、“Co-Parent ページ ID”、および、“リンクの重み”が、それぞれ関連付けられて記憶されている。

【0130】

リンク関係情報生成部251は、各々のページIDに対して、Co-Parentの関係にあるページIDを抽出するといった処理を行うとともに、Co-Parentページ間のリンクの重みも算出する。そのCo-Parentページ間のリンクの重みの算出は、以下のようにして行われる。すなわち、Co-Parentページ間のリンクの重み $W_o(u, v)$ は、次式(11)に基づいて算出される。

【0131】

$$W_o(u, v) = W_c(u, w) \times W_c(v, w) \cdots (11)$$

式(11)において、 u, v, w は、 P をページ集合とした場合、 $u, v, w \in P$ を満たす値であり、 $W_o(u, v)$ は、 $1 \leq W(u, v)$ を満たす値である。

【0132】

式(11)において、 $W_o(u, v)$ は、所定のページ u と、そのページ u とCo-Parentの関係にあるCo-Parentページ v 間のリンクの重みであり、 $W_c(u, w)$ は、所定のページ u と、そのページ u と親子関係にある子ページ w 間のリンクの重みであり、 $W_c(v, w)$ は、所定のページ v と、そのページ v と親子関係にある子ページ w 間のリンクの重みである。

【0133】

このようにして、Co-Parentページ間のリンクの重みが算出され、その算出結果が、図28に示したようなリスト形式のデータに書き込まれる。

【0134】

図27のフローチャートの説明に戻り、ステップS78において、CDFデータの生成が、CDFデータ生成部252(図26)により行われる。CDFデータ生成部252は、必要に応じ、リンク関係情報記憶部251(図26)と基本ページモデル記憶部163(図6)からデータを読み出し、その読み出したデータを用いて、図29に示すようなリスト形式のデータを作成し、CDFデータ記憶部262に記憶させる。

【0135】

ここで、CDFとは、Co-Parent Document Frequencyの略であり、各々のページに含まれる(各ページで出現する)単語に対して、その単語が現れる共通親ペ

ージのリンクの重みを総和したデータである。

【0136】

図29に示したCDFデータ記憶部262に記憶されるデータは、“ページID”と、“ページIDに含まれる単語IDと、その単語IDを含むCo-Parentページのリンクの重みの総和”といった情報を含む。このデータは、各々のページにおいて、そのページ内で出現する単語に対して、その単語が現れるCo-Parentページのリンクの重みを総和したデータであり、リンク判定部146が、前述のとおり $W_c(p, q) = 1$ とリンクの重みを生成した場合には、単にその単語が現れるCo-Parentページの総数となる。

【0137】

ステップS79において、ページモデル拡張部253（図26）は、ページモデル拡張処理を実行する。ページモデル拡張処理とは、図22に示すようなリスト形式のデータを作成し、ページモデル拡張データ記憶部263に記憶させる処理である。ページモデル拡張部253は、図22に示すようなデータを作成するために、基本ページモデル記憶部163、リンク情報記憶部164、リンク関係情報記憶部261、および、CDFデータ記憶部262に記憶されているデータを必要に応じて読み出す。

【0138】

図22に示したページモデル拡張データ記憶部263に記憶されているデータは、既に説明したように、“ページID”と“ベクトル”といった情報を含む。既に説明した実施の形態においては、Siblingの関係に注目したときのデータであったが、この実施の形態においては、Co-Parentの関係に注目したときのデータである。従って、そのデータの算出（“ベクトル”という情報内の“重み”という情報）に用いられる式が異なる。その異なる式に関して説明する。

【0139】

基本的に、Co-Parentの関係に注目し、ICDF（Inverse Co-Parent Document Frequency）に基づいて重みを計算した場合でも、“ベクトル”の重みに関するデータは、式（4）に基づいて算出される。ただし、式（4）に含まれる W_{ij} は、次式（12）に基づいて算出される。

$$W_{ij} = (1 + \log(TF_{ij})) \times (1 + \log(1 / (1 + CDF_{ij}))) \cdots (12)$$

式(12)において、 TF_{ij} は、単語jのページiにおける出現回数を示し、 $0 \leq TF_{ij}$ の値を取る。 CDF_{ij} は、ページiのCo-Parentページのうち、単語jを含むページのリンクの重みの総和を示す。

【0140】

このような式(4)と式(12)を用いてベクトル内の重みを算出するようにしても良いが、さらに、 CDF_{ij} の効果を高めるため、式(12)を式(13)に置き換えても良い。

$$W_{ij} = (1 + \log(TF_{ij})) \times (1 + \log(1 + ACDF_i / (1 + CDF_{ij}))) \cdots (13)$$

【0141】

式(13)において、 $ACDF_i$ は、ページiと全Co-Parentページ間のリンクの重みの総和を示す。

【0142】

さらに、 TTF_{ij} 、 ATF_{ij} を加え、式(12)を基に次式(14)あるいは、式(13)を基に次式(15)に基づいて、重みを算出するようにしても良い。

$$W_{ij} = (1 + \log(TF_{ij} + TTF_{ij} + ATF_{ij})) \times (1 + \log(1 / (1 + CDF_{ij}))) \cdots (14)$$

$$W_{ij} = (1 + \log(TF_{ij} + TTF_{ij} + ATF_{ij})) \times (1 + \log(1 + ACDF_i / (1 + CDF_{ij}))) \cdots (15)$$

式(14)、(15)において、 TTF_{ij} は、タグ付単語jのページiにおいて出現するか否かを示し、出現しない場合0が、出現する場合1が、それぞれ設定される。あるいは出現回数(0以上)が設定されるようにしても良い。タグの種類に応じてそれぞれ重みを付けるようにしても良い。

【0143】

また、 ATF_{ij} は、単語jのページiへのリンク元ページにおけるアンカー窓内において単語jが出現するか否かを示し、出現しない場合0が、出現する場合1が、それぞれ設定される。あるいは出現回数(0以上)が設定されるようにしても良い。タグ付単語と同様に、重み付けを行うようにしてもよい。さらに、アンカーか

らの距離に応じたウインドウ重みをつけるようにしても良い。

【0144】

ステップS80において、関連度算出部254において、ページ間の関連度が算出される。関連度算出部254は、必要に応じ、ページモデル拡張データ記憶部263に記憶されているデータを読み出し、図23に示すようなリスト形式のデータを作成し、関連度データ記憶部264に記憶させる。

【0145】

図23に示した関連度データ記憶部264に記憶されるデータは、既に説明したように、“ページID”、“対象ページID”、“関連度”、および、“高関連度単語”といった情報を含む。このうち、関連度は、Co-Parentの関係に注目して処理が行われる際でも、Siblingの関係に注目して処理が行われる際と同様の式により行われる。すなわち、既に説明した式(9)に基づいて算出される。

【0146】

ステップS81以降の処理は、図11のステップS51以降の処理と同様であるので、その説明は省略する。

【0147】

このように、Co-Parentの関係に注目して処理を行う場合においても、Siblingの関係に注目して処理を行う場合と同様の効果、又は、それ以上の効果を得ることが可能である。

【0148】

さらに、第3の実施の形態として、Siblingの関係とCo-Parentの関係の両方を考慮して処理を行うことが考えられる。そのようにした場合においても、検索サーバ4の構成は、図5乃至図7に示したような構成でよい。ただし、図7(図26)に示した詳細な構成は、図30に示したような構成とする。

【0149】

図30に示した検索サーバ4に含まれる内部構成例について、既に説明した図7又は図26と比較して説明する。図7に示したリンク関係情報生成部181または図26に示したリンク関係情報生成部251は、Siblingリンク関係情報生成部301とCo-Parentリンク関係情報生成部302で構成される。またこれら

の各部で生成されたデータを記憶するために、関連ページデータ記憶部 1 0 5 には、Siblingリンク関係情報記憶部 3 1 1 とCo-Parentリンク関係情報記憶部 3 1 2 とが、それぞれ設けられている。

【0 1 5 0】

図 7 に示した S D F データ生成部 1 8 2 または図 2 6 に示した C D F データ生成部 2 5 2 は、S D F ・ C D F データ生成部 3 0 3 で構成される。また、図 7 に示したページモデル拡張部 1 8 3 または図 2 6 に示したページモデル拡張部 2 5 3 は、I S D F ・ I C D F ページモデル拡張部 3 0 4 で構成される。これらの各部で生成されたデータを記憶するために、関連ページデータ記憶部 1 0 5 には、S D F ・ C D F データ記憶部 3 1 3 と I S D F ・ I C D F ページモデル拡張データ記憶部 3 1 4 が、それぞれ設けられている。

【0 1 5 1】

その他の部分に関しては、基本的に、図 7（図 2 6）に示した構成と同様なので、その説明は省略する。

【0 1 5 2】

図 3 1 のフローチャートを参照して、図 3 0 に示した構成を含む検索サーバ 4 の動作について説明する。ステップ S 1 0 1 乃至 S 1 0 6 の処理は、図 1 1 に示したフローチャートのステップ S 4 1 乃至 S 4 6 の処理と同様の処理であるので、その説明は省略する。

【0 1 5 3】

ステップ S 1 0 1 乃至 S 1 0 6 における処理、すなわち、検索サーバ 4 内の構成のうち、図 6 に示した部分で行われる処理が行われることにより、図 6 に示した、保存ページ記憶部 1 0 2、ページ I D 記憶部 1 6 1、単語 I D 記憶部 1 6 2、基本ページモデル記憶部 1 6 3、および、リンク情報記憶部 1 6 4 にはそれぞれ、図 1 4 乃至図 1 8 に示したデータが記憶される。

【0 1 5 4】

ステップ S 1 0 7 において、Siblingリンク関係情報が、Siblingリンク関係情報生成部 3 0 1（図 3 0）により生成されるわけだが、その生成され、Siblingリンク関係情報記憶部 3 1 1 に記憶されるデータは、図 1 9 に示したようなデー

タである。すなわち、ステップS107における処理は、図11のステップS47の処理と同様であり、Siblingリンク関係情報生成部301が生成するデータは、図7に示したリンク情報関係情報生成部181が生成するデータと同様であるので、その詳細な説明は既に説明したので、ここではその説明を省略する。

【0155】

次に、ステップS108において、Co-ParentTリンク関係情報が、Co-Parentリンク関係情報生成部302により生成されるわけだが、その生成され、Co-Parentリンク関係情報記憶部312に記憶されるデータは、図28に示したようなデータである。すなわち、ステップS108における処理は、図27のステップS77の処理と同様であり、Co-Parentリンク関係情報生成部302が生成するデータは、図26に示したリンク情報関係情報生成部251が生成するデータと同様であるので、その詳細な説明は既に説明したので、ここではその説明を省略する。

【0156】

図31のフローチャートの説明に戻り、ステップS109において、SDF・CDFデータの生成が、SDF・CDFデータ生成部303（図30）により行われる。SDF・CDFデータ生成部303は、必要に応じ、Siblingリンク関係情報記憶部311、Co-Parentリンク関係情報記憶部312、および基本ページモデル記憶部163（図6）からデータを読み出し、その読み出したデータを用いて、図21と図29に示すようなリスト形式のデータを作成し、SDF・CDFデータ記憶部313に記憶させる。

【0157】

図21に示したデータは、SDF用のデータであり、図29に示したデータは、CDF用のデータである。SDF用のデータは、図7のSDFデータ生成部182が図11のステップS48の処理として行う処理と同様な処理により生成され、CDF用のデータは、図26のCDFデータ生成部252が図27のステップS78の処理として行う処理と同様な処理により生成される。これらの生成については、既に説明したので、ここでは、その説明を省略する。

【0158】

また、図 21 と図 29 に示したリスト形式のデータは、それぞれ別々のリスト形式のデータとして、SDF・CDF データ記憶部 313 に記憶されるようにしても良いし、1つのリスト形式としてまとめられて記憶されるようにしても良い。

【0159】

ステップ S110 において、ISDF・ICDF ページモデル拡張部 304 (図 30) は、ISDF・ICDF ページモデル拡張処理を実行する。ISDF・ICDF ページモデル拡張処理とは、図 22 に示すようなリスト形式のデータを作成し、ISDF・ICDF ページモデル拡張データ記憶部 314 に記憶させる処理である。

【0160】

ISDF・ICDF ページモデル拡張データ記憶部 314 に記憶されているデータは、図 22 に示したようなデータであるとし、その図 22 に示したデータは、既に説明したように、“ページ ID” と “ベクトル” といった情報を含む。図 22 に示したデータについては、Sibling の関係に注目したときのデータ、または、Co-Parent の関係に注目したときのデータであるとして説明した。ここでは、その両方の関係に注目したときのデータであるため、そのデータの算出 (“ベクトル” という情報内の “重み” という情報) に用いられる式が異なる。その異なる式に関して説明する。

【0161】

基本的に、Sibling の関係と Co-Parent の関係の両方に注目したときでも、“ベクトル” の重みに関するデータは、式 (4) に基づいて算出される。式 (4) に含まれる W_{ij} は、次式 (16) に基づいて算出される。

$$W_{ij} = (1 + \log(TF_{ij})) \times (1 + \log(1 / (1 + SDF_{ij} + CDF_{ij}))) \quad \dots (16)$$

式 (16) において、 TF_{ij} は、単語 j のページ i における出現回数を示し、 $0 \leq TF_{ij}$ の値を取る。 SDF_{ij} は、ページ i の Sibling ページのうち、単語 j を含むページのリンクの重みの総和を示し、 CDF_{ij} は、ページ i の Co-Parent ページのうち、単語 j を含むページのリンクの重みの総和を示す。

【0162】

このような式(4)と式(16)を用いてベクトル内の重みを算出するようにしても良いが、さらに、SDFijとCDFijの効果を、それぞれ高めるため、式(16)を式(17)に置き換えて算出するようにしても良い。

$$W_{ij} = (1 + \log(TF_{ij})) \times (1 + \log(1 + ACDF_i + ASDFi / (1 + ASDF_{ij} + CDF_{ij}))) \cdots (17)$$

【0163】

式(17)において、ASDFiは、ページiと全Siblingページ間のリンクの重みの総和を、ACDFiは、ページiと全Co-Parentページ間のリンクの重みの総和を示す。

【0164】

さらに、TTFij, ATFijを加え、式(16)を基に次式(18)あるいは、式(17)を基に次式(19)に基づいて、重みを算出するようにしても良い。

$$W_{ij} = (1 + \log(TF_{ij} + TTF_{ij} + ATF_{ij})) \times (1 + \log(1 / (1 + SDF_{ij} + CDF_{ij}))) \cdots (18)$$

$$W_{ij} = (1 + \log(TF_{ij} + TTF_{ij} + ATF_{ij})) \times (1 + \log(1 + ASDFi + ACDF_i / (1 + SDF_{ij} + CDF_{ij}))) \cdots (19)$$

式(18)または式(19)において、TTFijは、タグ付単語jのページiにおいて出現するか否かを示し、出現しない場合0が、出現する場合1が、それぞれ設定される。あるいは出現回数(0以上)が設定されるようにしても良い。タグの種類に応じてそれぞれ重みを付けるようにしても良い。

【0165】

また、ATFijは、単語jのページiへのリンク元ページにおけるアンカー窓内において単語jが出現するか否かを示し、出現しない場合0が、出現する場合1が、それぞれ設定される。あるいは出現回数(0以上)が設定されるようにしても良い。タグ付単語と同様に、重み付けを行うようにしてもよい。さらに、アンカーからの距離に応じたウインドウ重みをつけるようにしても良い。

【0166】

図31のフローチャートの説明に戻り、ステップS111において、関連度算

出部 305 において、ページ間の関連度が算出される。関連度算出部 305 は、必要に応じ、ISDF・ICDF ページモデル拡張データ記憶部 314 に記憶されているデータを読み出し、図 23 に示すようなリスト形式のデータを作成し、関連度データ記憶部 315 に記憶させる。

【0167】

図 23 に示した関連度データ記憶部 264 に記憶されるデータは、既に説明したように、“ページ ID”、“対象ページ ID”、“関連度”、および、“高関連度単語”といった情報を含む。このうち、関連度は、Co-Parent の関係に注目して処理が行われる際でも、Sibling の関係に注目して処理が行われる際でも、または、Sibling と Co-Parent の両方の関係に注目して処理が行われる際でも、同様の式により行われる。すなわち、既に説明した式 (9) に基づいて算出される。

【0168】

ステップ S112 以降の処理は、図 11 のステップ S51 以降の処理と同様であるので、その説明は省略する。

【0169】

このように、Sibling の関係と Co-Parent の関係の両方に注目して処理を行う場合においても、Co-Parent の関係に注目して処理を行うときや、Sibling の関係に注目して処理を行うときと同様の効果、またはそれ以上の効果を得ることが可能である。

【0170】

上述した実施の形態においては、ユーザに関連ページの情報を提供する際の処理について説明したが、その関連ページの情報に、広告などの情報を含めるようにしても良い。そのような広告などの情報も提供するようにした場合、検索サーバ 4 の構成は、図 32 に示したようになる。図 32 に示した検索サーバ 4 の構成は、図 5 に示した検索サーバ 4 の構成に、特殊設定管理用記憶部 331 を追加した構成とされている。

【0171】

この特殊設定管理用記憶部 331 には、図 33、図 34 にそれぞれ示す記憶部

が設けられている。図 33 に示した特殊設定用管理データ記憶部 341 には、“タイトル”、“リンク先 URL”、“説明”、“単語”、“URL パターン”、および、“オーナ ID”といった情報が含まれている。図 34 に示した特殊設定管理者データ記憶部 342 には、“オーナ ID”、“名前”、“所属”、“e-mail”、“Account”、および、“Password”といった情報が含まれている。

【0172】

このような特殊設定管理用記憶部 331 が、検索サーバ 4 に設けられた場合、例えば、図 11 に示したフローチャートにおいて、関連ページ一覧生成という処理の内の 1 処理として、この特殊設定管理用記憶部 331 に記憶されている情報を提供するための処理が実行される。具体的には、関連ページの一覧表のデータが、作成された後に、特殊設定管理用記憶部 331 が参照され、その関連ページに関連すると判断される URL などの情報が、特殊設定用管理データ記憶部 341 から抽出され、一覧表のデータに含まれる。

【0173】

提供されたデータがユーザ側の端末 3 で再生されると、その画面には、関連ページの一覧と、その関連ページに関わりのある情報（広告）が表示されている。

【0174】

特殊設定用管理データ記憶部 341 に記憶されているデータは、管理者により削除、追加、訂正などの処理が行えるようになっており、その管理者を管理するためのデータが、特殊設定管理者データ記憶部 342 に記憶されている。この特殊設定管理者データ記憶部 342 に記憶されている管理者のみが、特殊設定用管理データ記憶部 341 のデータを操作することが可能とするために、パスワード (Password) などが設定されるようになっている。

【0175】

このように、関連ページの一覧表に、広告も含めるようにした場合、その広告を掲載する会社から、その掲載料金を徴収することが可能となる。また、上述した実施の形態においては説明しなかったが、例えば、検索サーバ 4 の収集サイトリスト記憶部 101 に記憶されるサイトを管理する管理者から、料金を徴収するようにしても良い。

【 0 1 7 6 】

これは、検索サーバ 4 により、ユーザに関連ページであるとしてユーザに提供されることにより、そのサイトへのアクセスの増加を期待することができ、そのために、検索サーバ 4 自体に登録してもらいたいというサイトの管理者から登録料として料金を徴収することができる。

【 0 1 7 7 】

このような課金制度を、必要に応じて設けることも可能である。

【 0 1 7 8 】

上述した一連の処理は、それぞれの機能を有するハードウェアにより実行させることもできるが、ソフトウェアにより実行させることもできる。一連の処理をソフトウェアにより実行させる場合には、そのソフトウェアを構成するプログラムが専用のハードウェアに組み込まれているコンピュータ、または、各種のプログラムをインストールすることで、各種の機能を実行することが可能な、例えば汎用のパーソナルコンピュータなどに、記録媒体からインストールされる。

【 0 1 7 9 】

記録媒体は、図 2 に示すように、WWWサーバ 2 を構成するパーソナルコンピュータとは別に、ユーザにプログラムを提供するために配布される、プログラムが記録されている磁気ディスク 3 1（フレキシブルディスクを含む）、光ディスク 3 2（CD-ROM（Compact Disc-Read Only Memory）、DVD（Digital Versatile Disc）を含む）、光磁気ディスク 3 3（MD（Mini-Disc）（登録商標）を含む）、若しくは半導体メモリ 3 4 などよりなるパッケージメディアにより構成されるだけでなく、コンピュータに予め組み込まれた状態でユーザに提供される、プログラムが記憶されている ROM 1 2 や記憶部 1 8 が含まれるハードディスクなどで構成される。

【 0 1 8 0 】

なお、本明細書において、媒体により提供されるプログラムを記述するステップは、記載された順序に従って、時系列的に行われる処理は勿論、必ずしも時系列的に処理されなくとも、並列的あるいは個別に実行される処理をも含むものである。

【 0 1 8 1 】

また、本明細書において、システムとは、複数の装置により構成される装置全体を表すものである。

【 0 1 8 2 】**【発明の効果】**

本発明の情報処理装置および方法、記録媒体、並びにプログラムによれば、インターネット上に開設されているサイトの検索を行うことが可能である。

【 0 1 8 3 】

また、本発明の情報処理装置および方法、記録媒体、並びにプログラムによれば、よりユーザの所望としているサイトを検索し、その情報を提供することが可能である。

【図面の簡単な説明】**【図 1】**

本発明を適用した情報処理システムの一実施の形態の構成を示す図である。

【図 2】

WWWサーバの内部構成例を示す図である。

【図 3】

端末 3 の内部構成例を示す図である。

【図 4】

検索サーバの内部構成例を示す図である。

【図 5】

検索サーバの内部構成例を示す図である。

【図 6】

検索サーバの詳細な内部構成例を示す図である。

【図 7】

検索サーバの詳細な内部構成例を示す図である。

【図 8】

リンク関係について説明するための図である。

【図 9】

端末と検索サーバとの間で行われる処理について説明するフローチャートである。

【図 1 0】

端末側のディスプレイ上に表示される画面の一例を示す図である。

【図 1 1】

検索サーバの動作について説明するためのフローチャートである。

【図 1 2】

収集サイトリスト記憶部に記憶されるデータを説明するための図である。

【図 1 3】

保存ページ記憶部に記憶されるサイトのデータを説明するための図である。

【図 1 4】

ページ I D 記憶部に記憶されるデータを説明するための図である。

【図 1 5】

単語 I D 記憶部に記憶されるデータを説明するための図である。

【図 1 6】

単語 I D 記憶部に記憶されるデータを説明するための図である。

【図 1 7】

基本ページモデル記憶部に記憶されるデータを説明するための図である。

【図 1 8】

リンク情報記憶部に記憶されるデータを説明するための図である。

【図 1 9】

リンク関係情報記憶部に記憶されるデータを説明するための図である。

【図 2 0】

重みの算出について説明するための図である。

【図 2 1】

S D F データ記憶部に記憶されているデータを説明するための図である。

【図 2 2】

ページモデル拡張データ記憶部に記憶されるデータを説明するための図である。

。

【図 2 3】

関連度データ記憶部に記憶されるデータを説明するための図である。

【図 2 4】

関連ページ間の特徴の抽出について説明するための図である。

【図 2 5】

リンク関係について説明するための図である。

【図 2 6】

検索サーバの詳細な他の内部構成例を示す図である。

【図 2 7】

図 2 6 に示した構成を有する検索サーバの動作について説明するフローチャートである。

【図 2 8】

リンク関係情報記憶部に記憶されるデータを説明するための図である。

【図 2 9】

C D F データ記憶部 2 6 2 に記憶されるデータを説明するための図である。

【図 3 0】

検索サーバの詳細な他の内部構成例を示す図である。

【図 3 1】

図 3 0 に示した構成を有する検索サーバの動作について説明するフローチャートである。

【図 3 2】

検索サーバの他の内部構成例を示す図である。

【図 3 3】

特殊設定用管理データ記憶部に記憶されるデータを説明する図である。

【図 3 4】

特殊設定管理者データ記憶部に記憶されるデータを説明する図である。

【符号の説明】

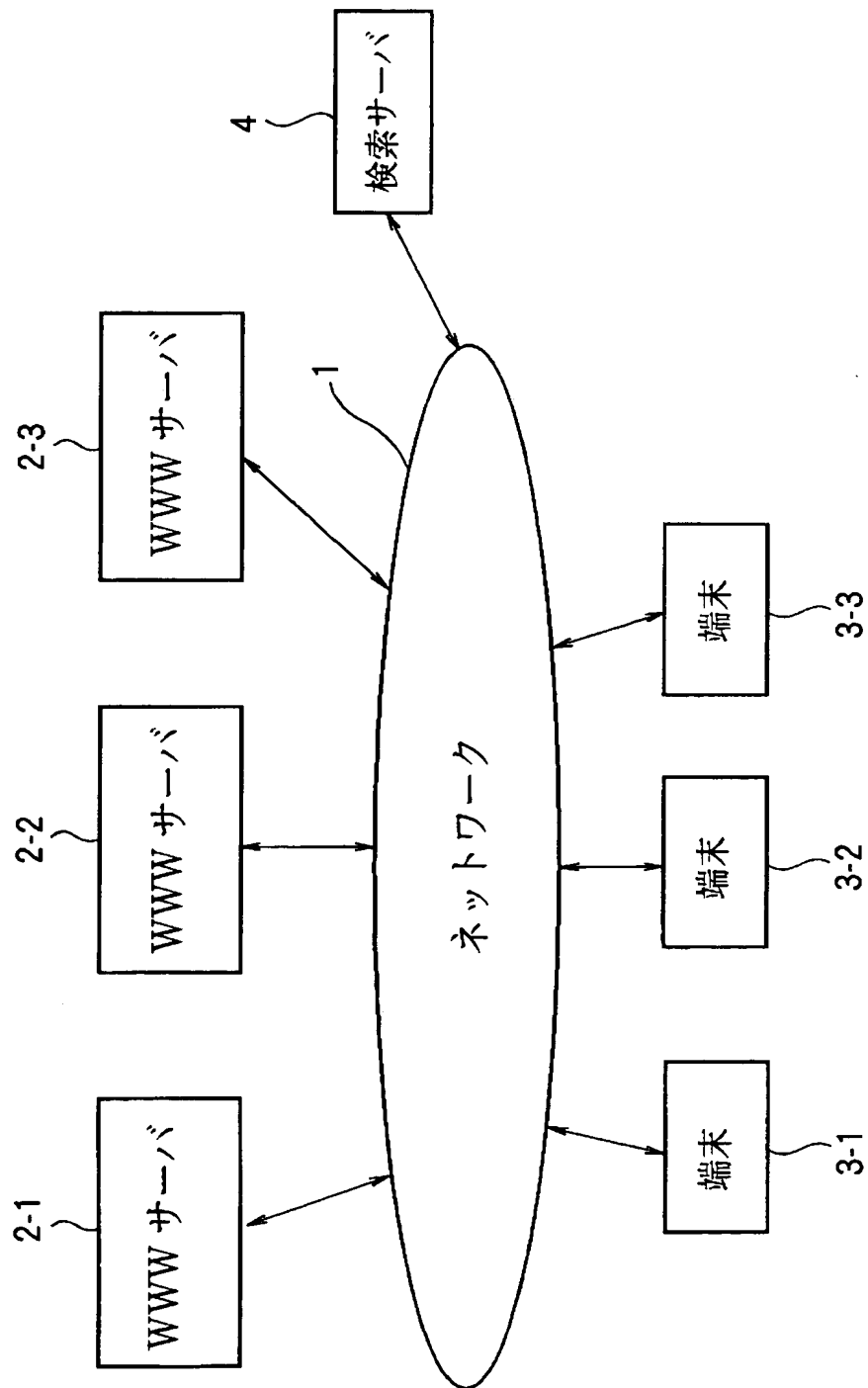
1 ネットワーク, 2 WWWサーバ, 3 端末, 4 検索サーバ,
1 0 1 収集サイトリスト記憶部, 1 0 2 保存ページ記憶部, 1 0 3 ペ

ージデータ記憶部, 1 0 4 サイトページデータ記憶部, 1 0 5 関連ページデータ記憶部, 1 1 1 サイトページ記憶部, 1 1 2 関連ページデータ処理部

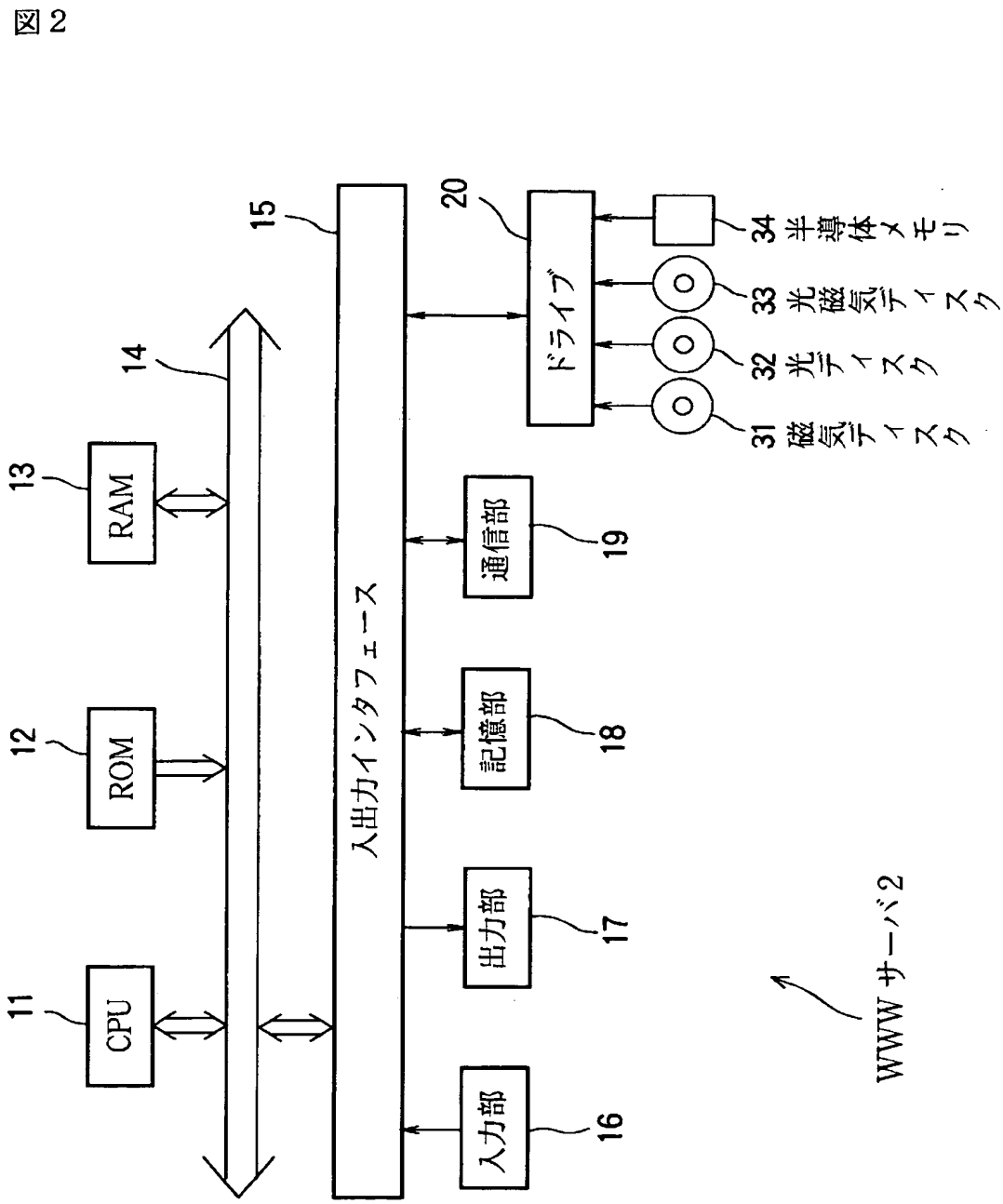
【書類名】 図面

【図 1】

図 1

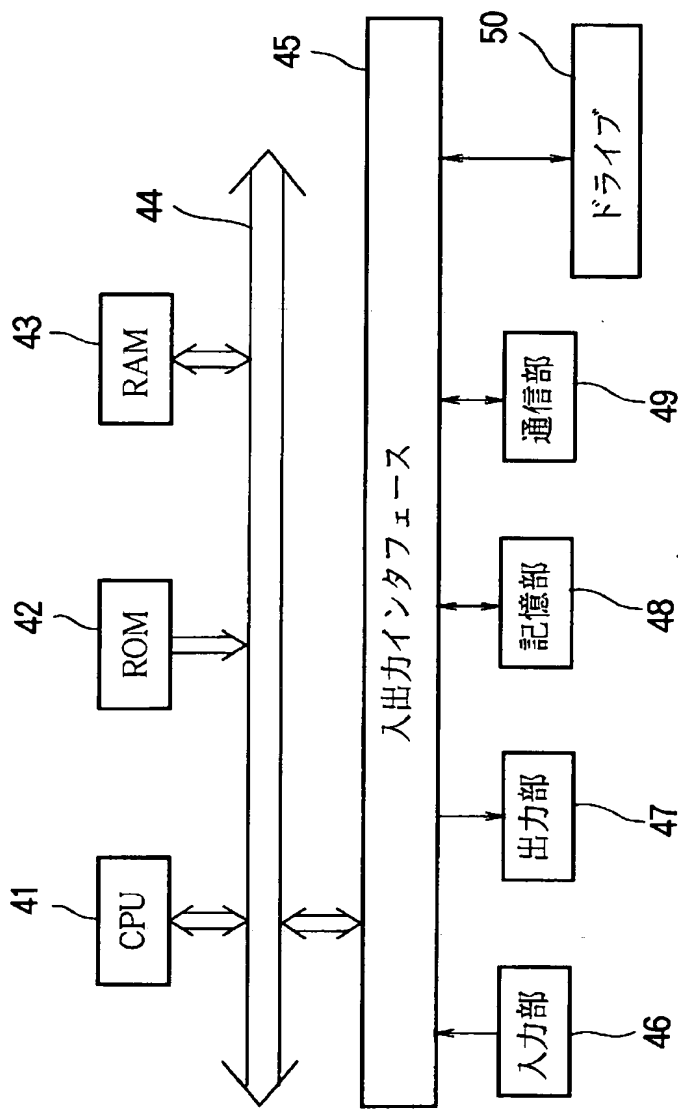


【図 2】



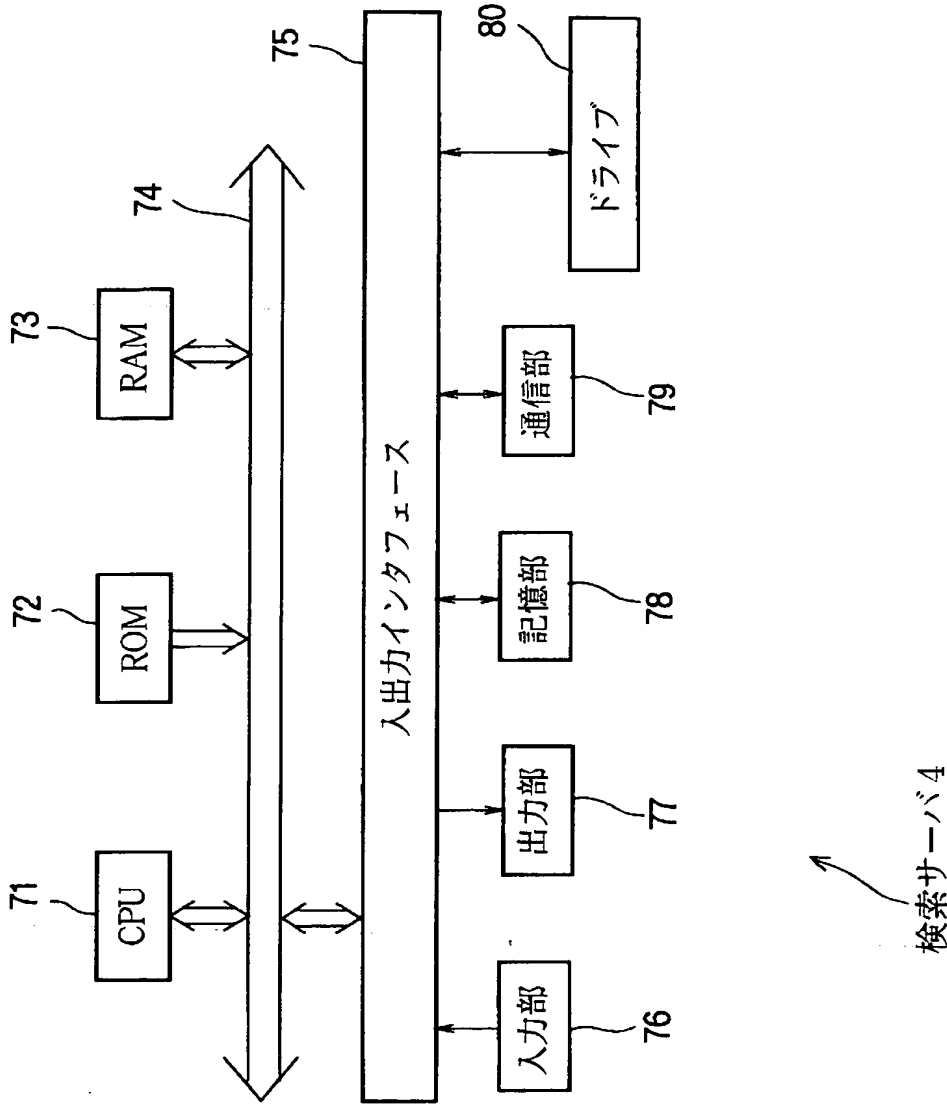
【図 3】

図 3



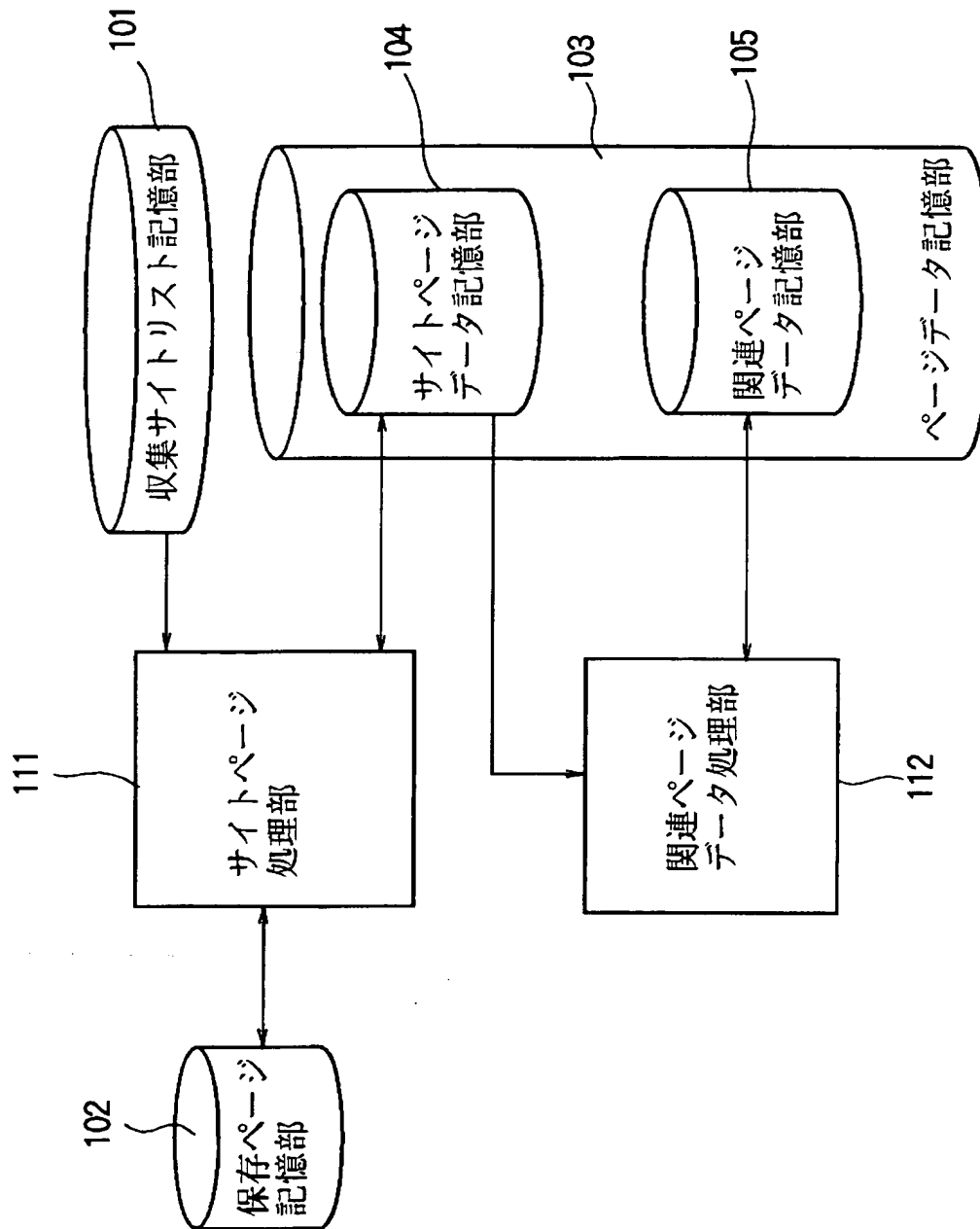
【図 4】

図 4

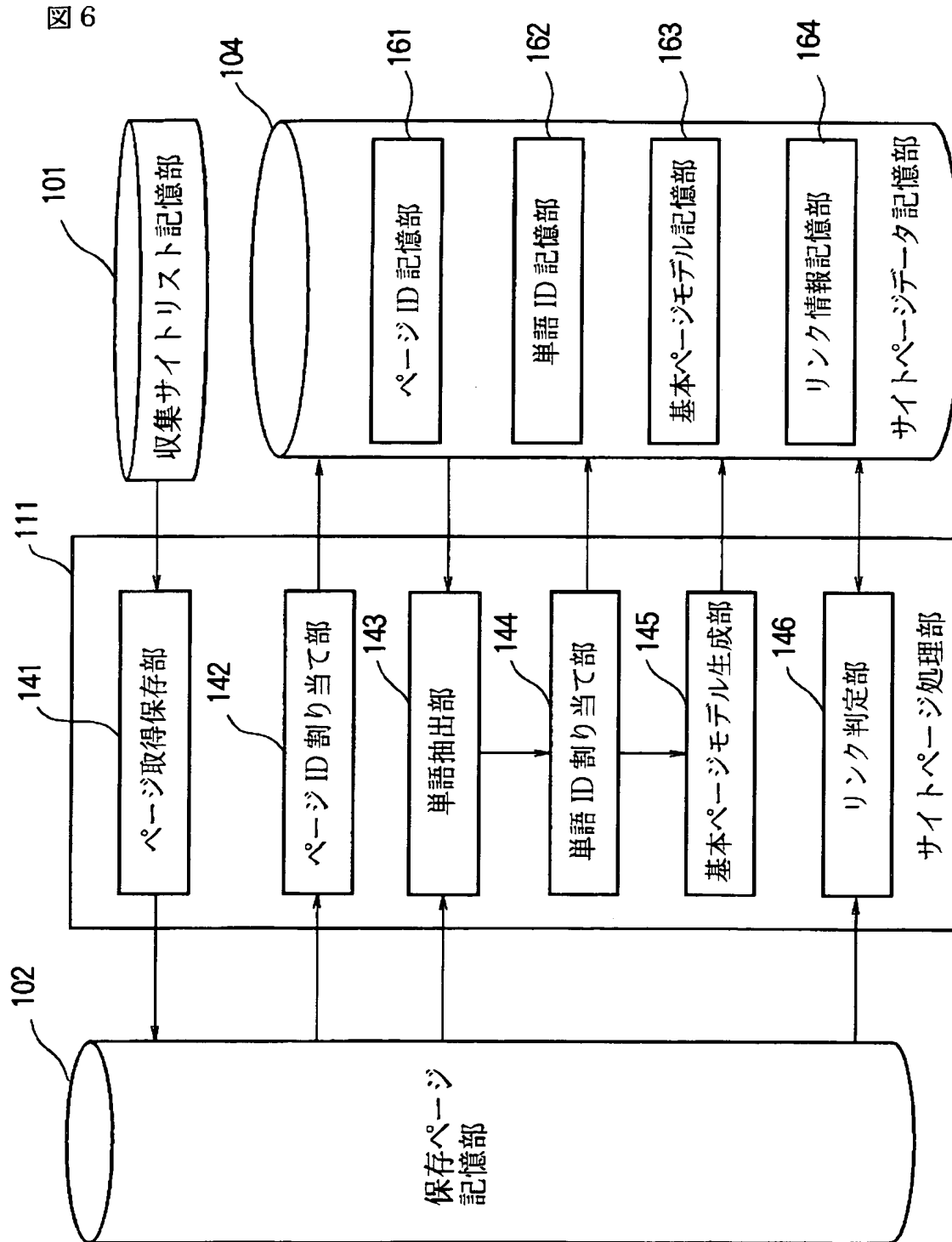


【図5】

図5

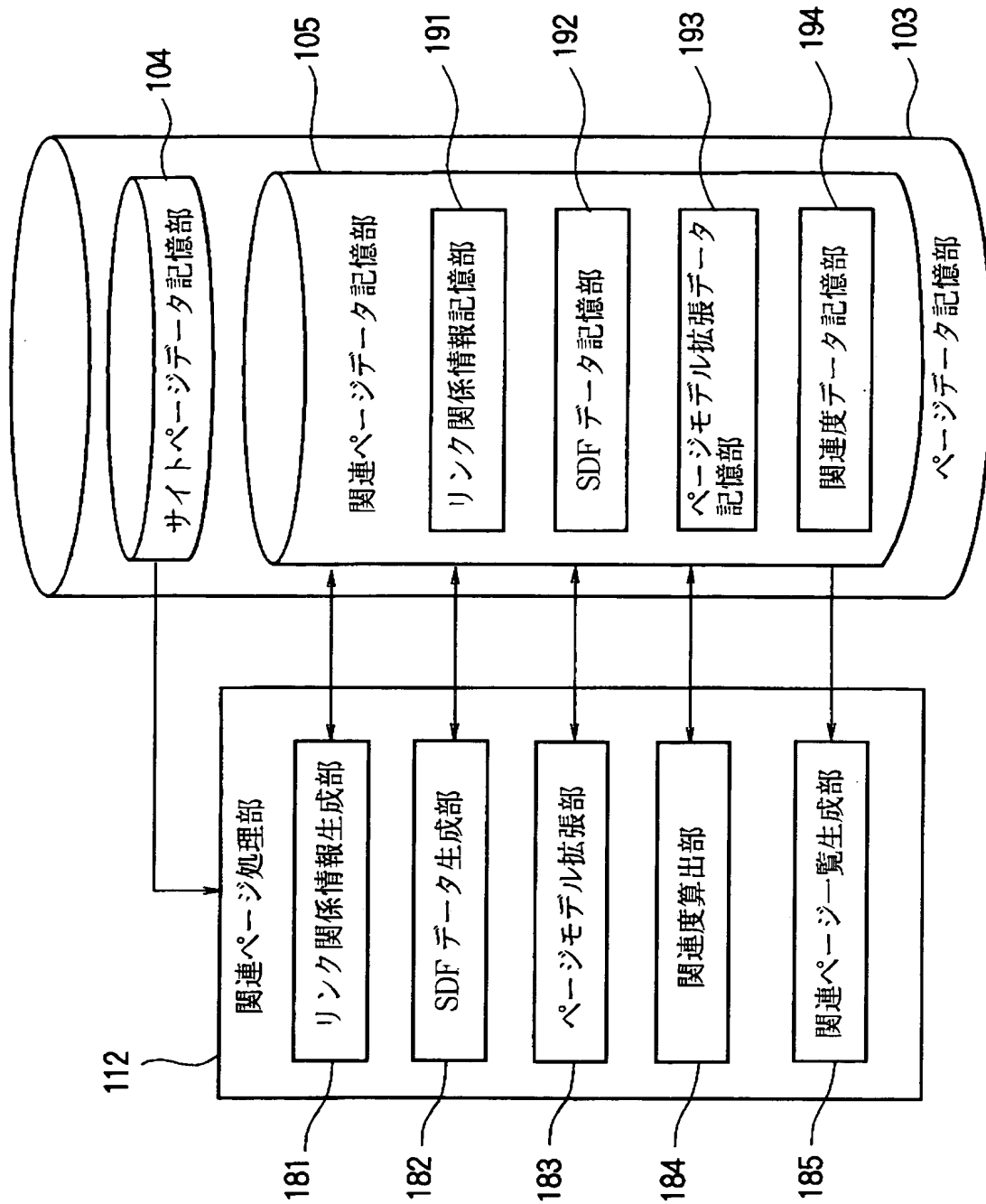


【図 6】



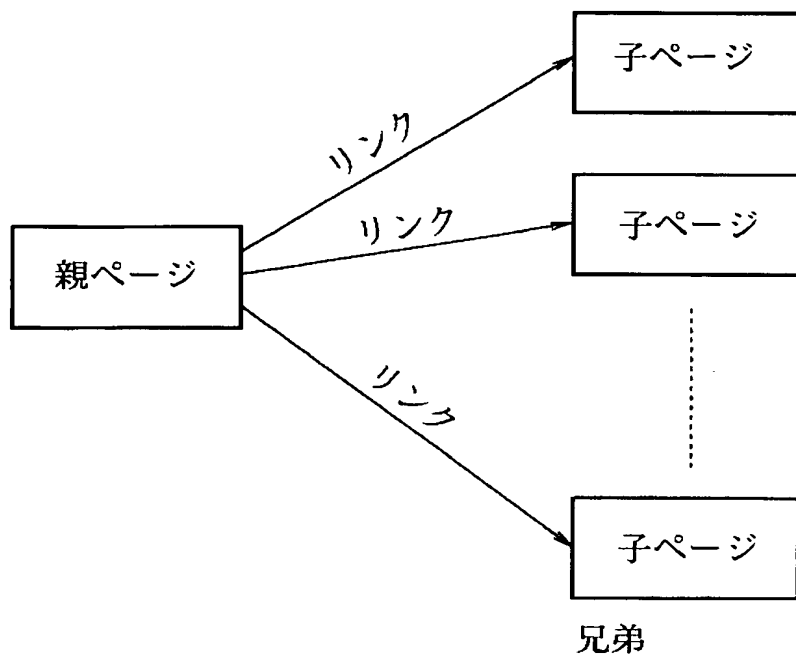
【図 7】

図 7



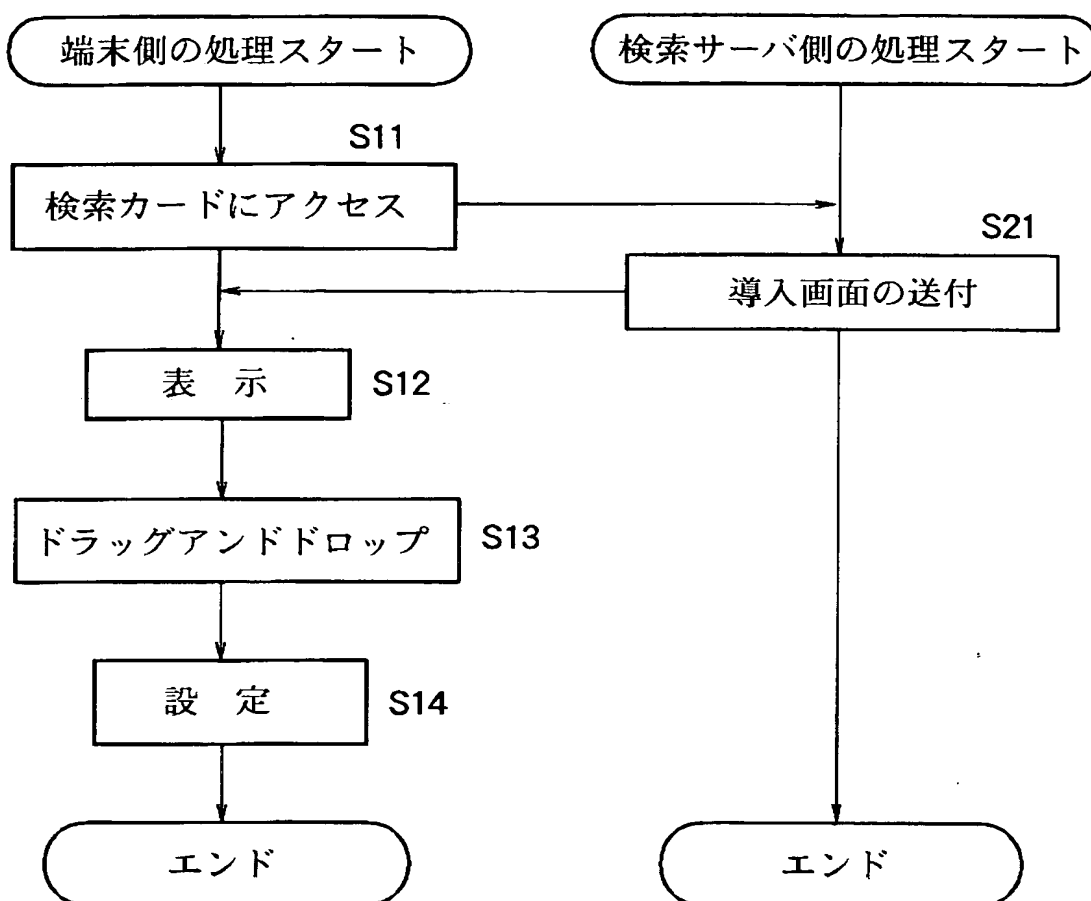
【図 8】

図 8



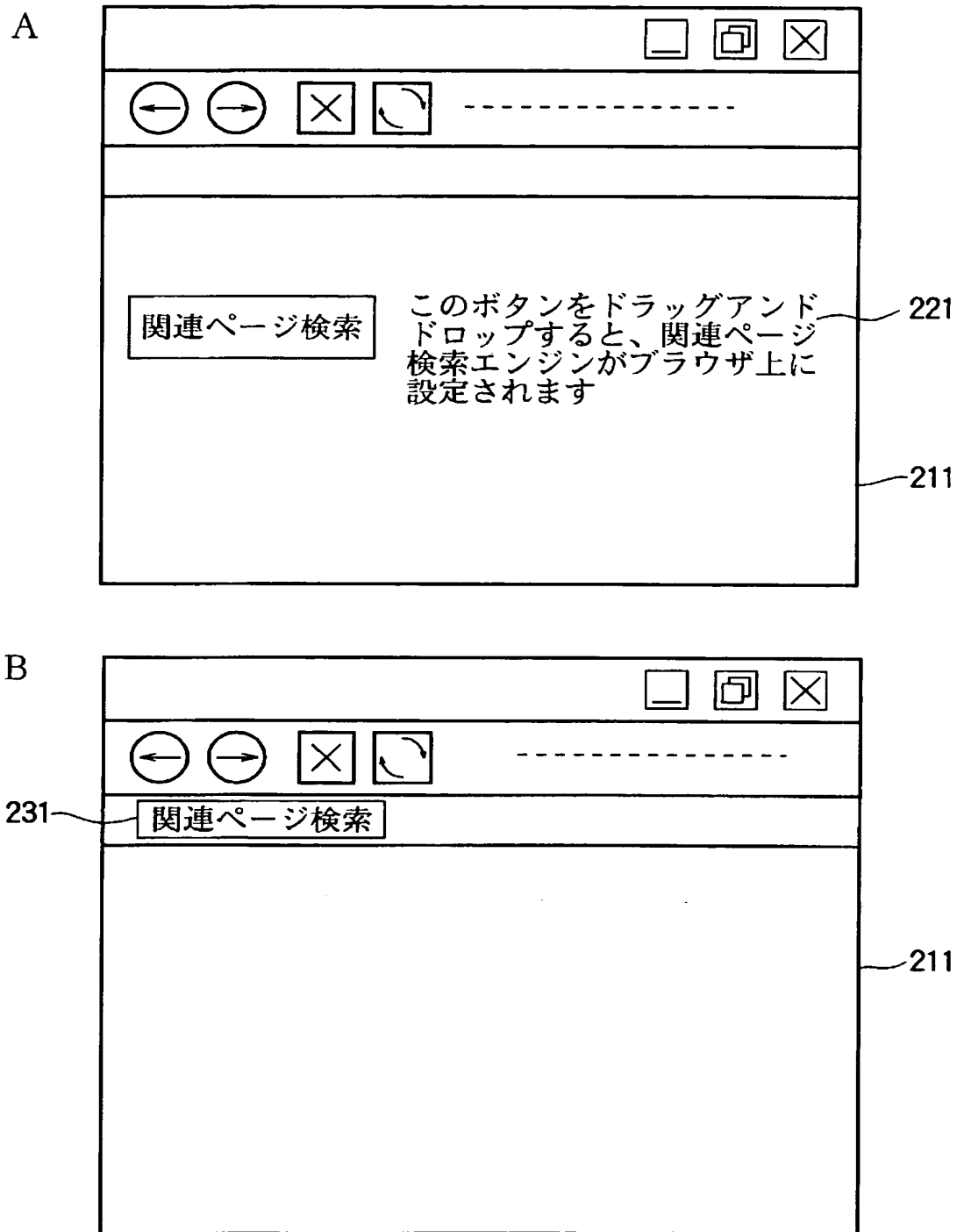
【図 9】

図 9



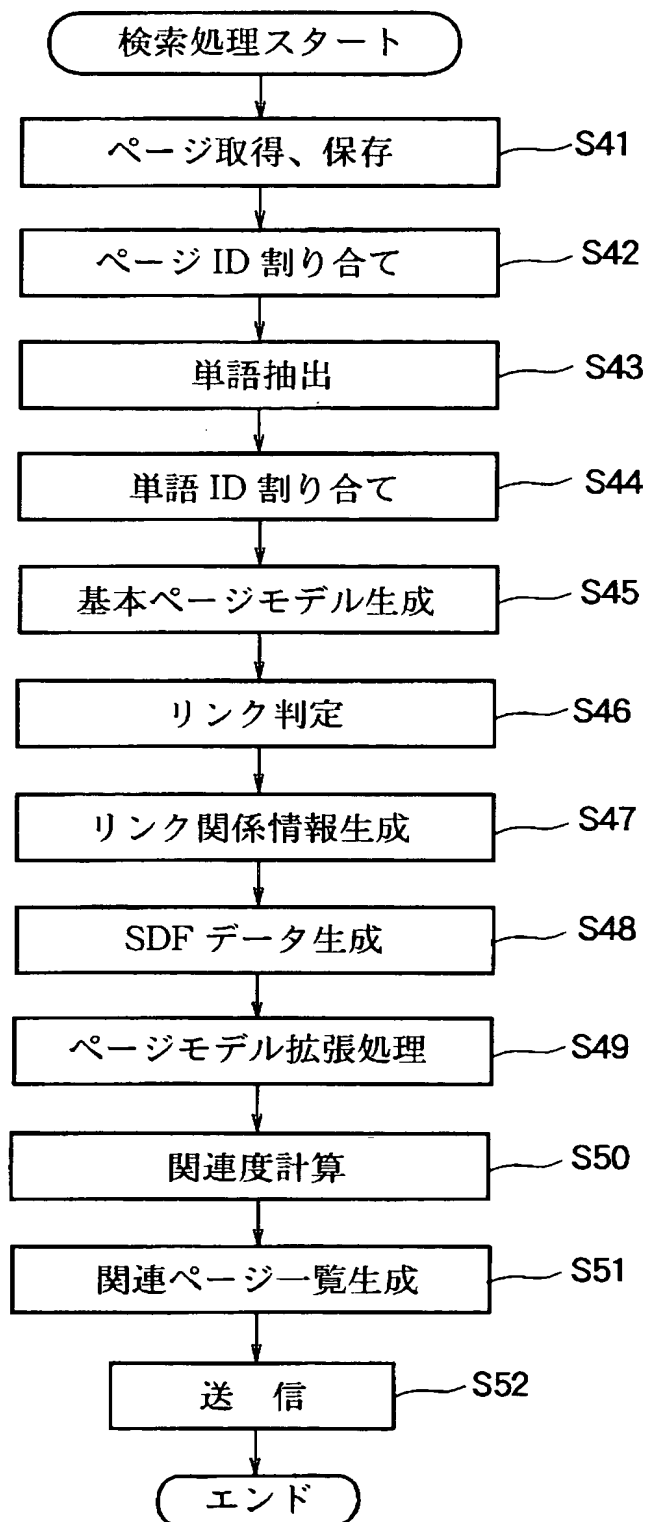
【図 10】

図 10



【図 11】

図 11



【図 1 2】

図 12

収集開始 URL (Key)	含む ディレクトリ	排他 ディレクトリ	含む ドメイン	排他 ドメイン
www.ssss.co.jp/index.html				
www.vaaa.sss.co.jp/index.html				
www.spe.co.jp/index.html				

101

【図 1 3】

図 13

サイト ID (Key)	サイト名	総ページ数
0001	www.ssss.co.jp	12345
0002	www.ssss.music.co.jp	4321

102

【図 1 4】

図 14

ページ ID (Key)	サイト ID	ページ URL	タイトル	サマリー	ページ保存場所	最終更新日
00010000001	0001	www.ssss.co.jp/i ndex.html	S s s s	Ssss Corp.	A/index.html	2002-07-10
00020000001	0002	www.spe.co.jp/in dex.html	S P E	SPE Japan	B/index.html	2002-07-10

161

【図 1 5】

図 15

単語 ID (Key)	単語
000001	映画
000002	MUSIC

162



【図 1 6】

図 16

単語 ID (Key1)	サイト ID (Key2)	そのサイト内で 当該単語を含む ページ数	そのサイト内で当該単語を含むページ ID
000001	001	30	0001000001,0001000002,.....
000001	002	15	0002000001,0002000004,.....

162

【図 1 7】

図 17

ページ ID (Key)	出現単語	Title	Keywords	description
0001000001	{単語 ID, 出現数}	{単語 ID, 出現数}	{単語 ID, 出現数}	{単語 ID, 出現数}
0001000002	{単語 ID, 出現数}	{単語 ID, 出現数}	{単語 ID, 出現数}	{単語 ID, 出現数}

163

【図 1 8】

図 18

ページ ID (Key1)	リンク先ページ ID (Key2)	リンクの重み	アンカー窓内単語
00010000001	00010000002	1.4	{単語 ID, 出現数}
00010000001	00010000005	1.3	{単語 ID, 出現数}

164

【図 1 9】

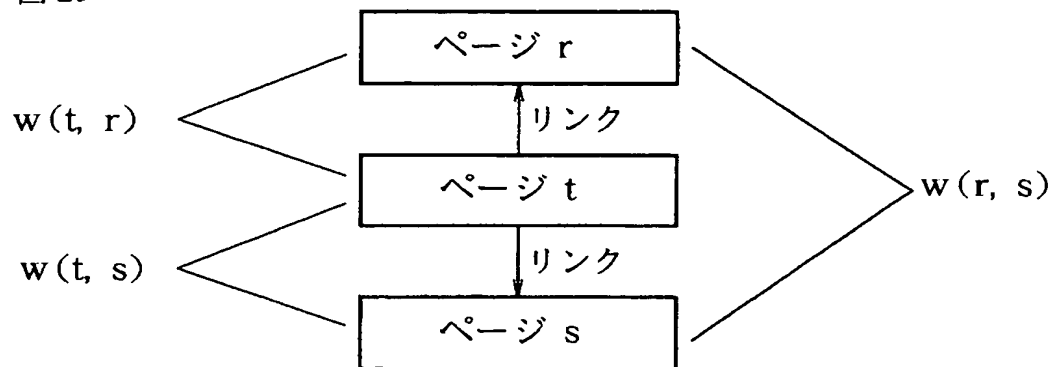
図 19

ページ ID (Key1)	Sibling ページ ID (Key2)	リンクの重み
000100000005	000100000007	1.5
000100000005	000200000002	1.3

191

【図 20】

図 20



【図 21】

図 21

ページ ID (Key)	ページ ID に含まれる単語 ID と、その単語 ID を含む Sibling ページのリンクの重みの総和
0001000001	{単語 ID, リンクの重みの総和}, {単語 ID, リンクの重みの総和},
0001000002	{単語 ID, リンクの重みの総和}, {単語 ID, リンクの重みの総和},

192

【図 22】

図 22

ページ ID (Key)	ベクトル
0001000001	{単語 ID, 重み},
0001000002	{単語 ID, 重み},

193

【図 2 3】

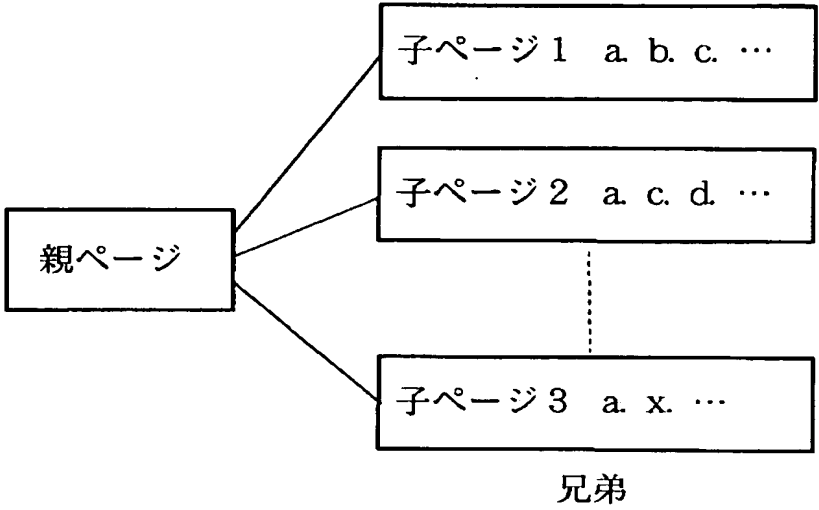
図 23

ページ ID (Key1)	対象ページ ID (Key2)	関連度	高関連度単語
0001000001	0001000002	0.10	{単語 ID, 単語関連度}
0001000001	0001000003	0.40	{単語 ID, 単語関連度}
0001000001	0001000004	0.12	{単語 ID, 単語関連度}

194

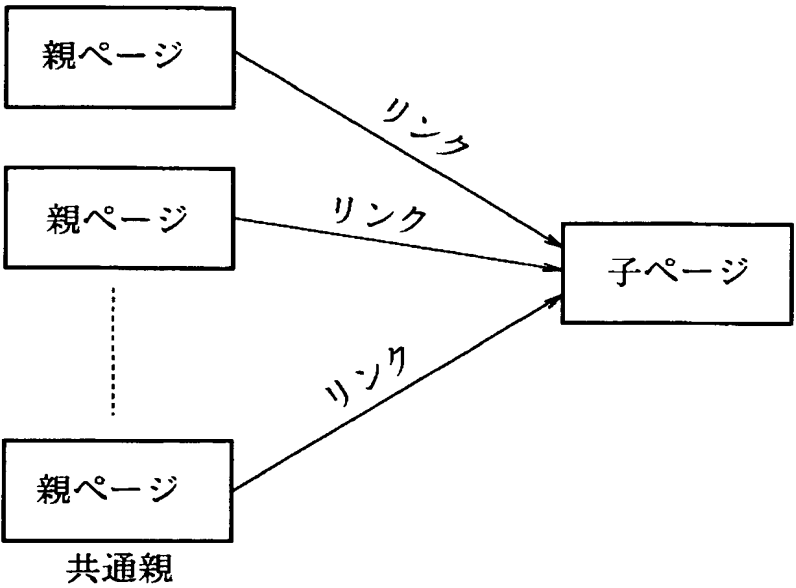
【図 24】

図 24

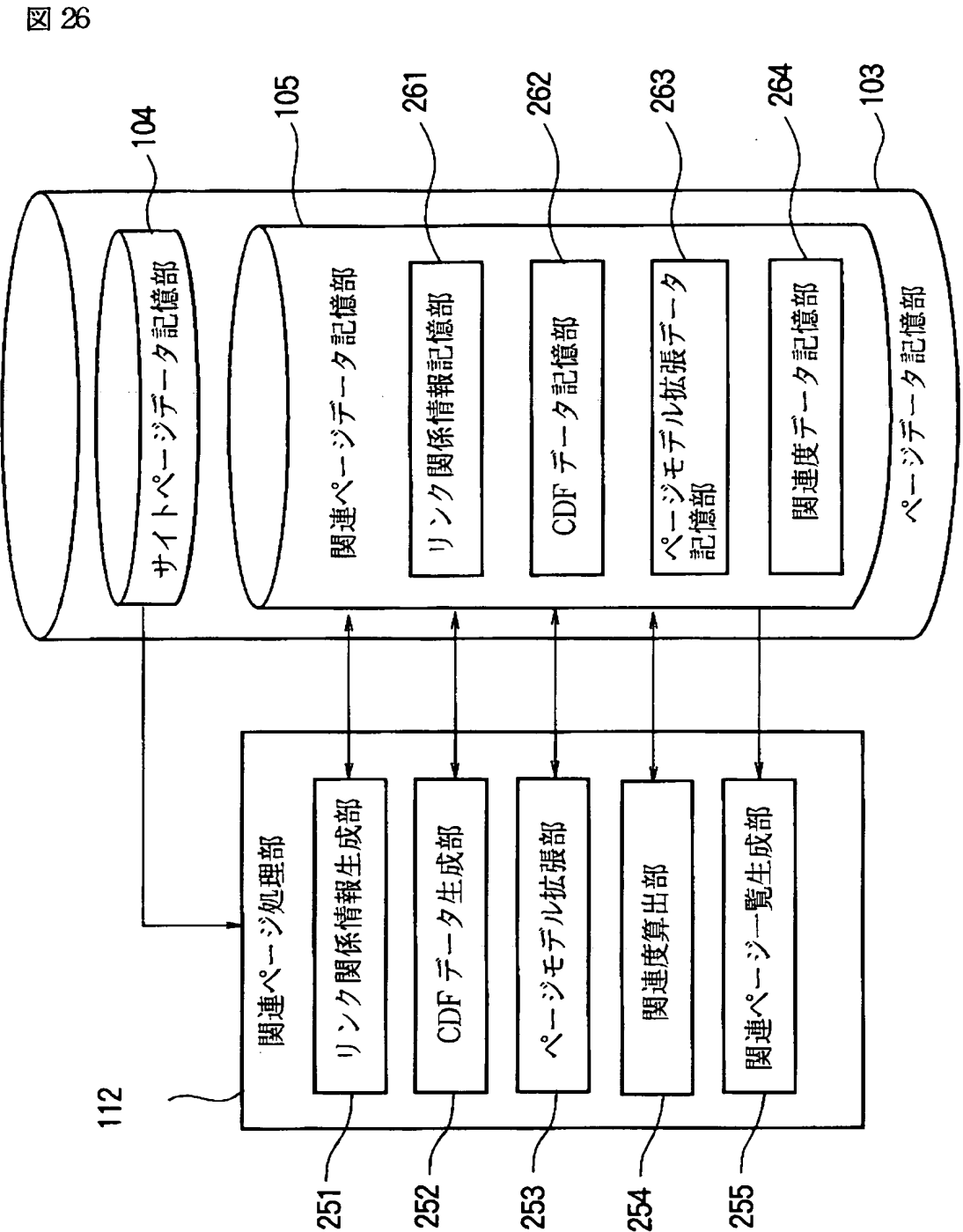


【図 25】

図 25

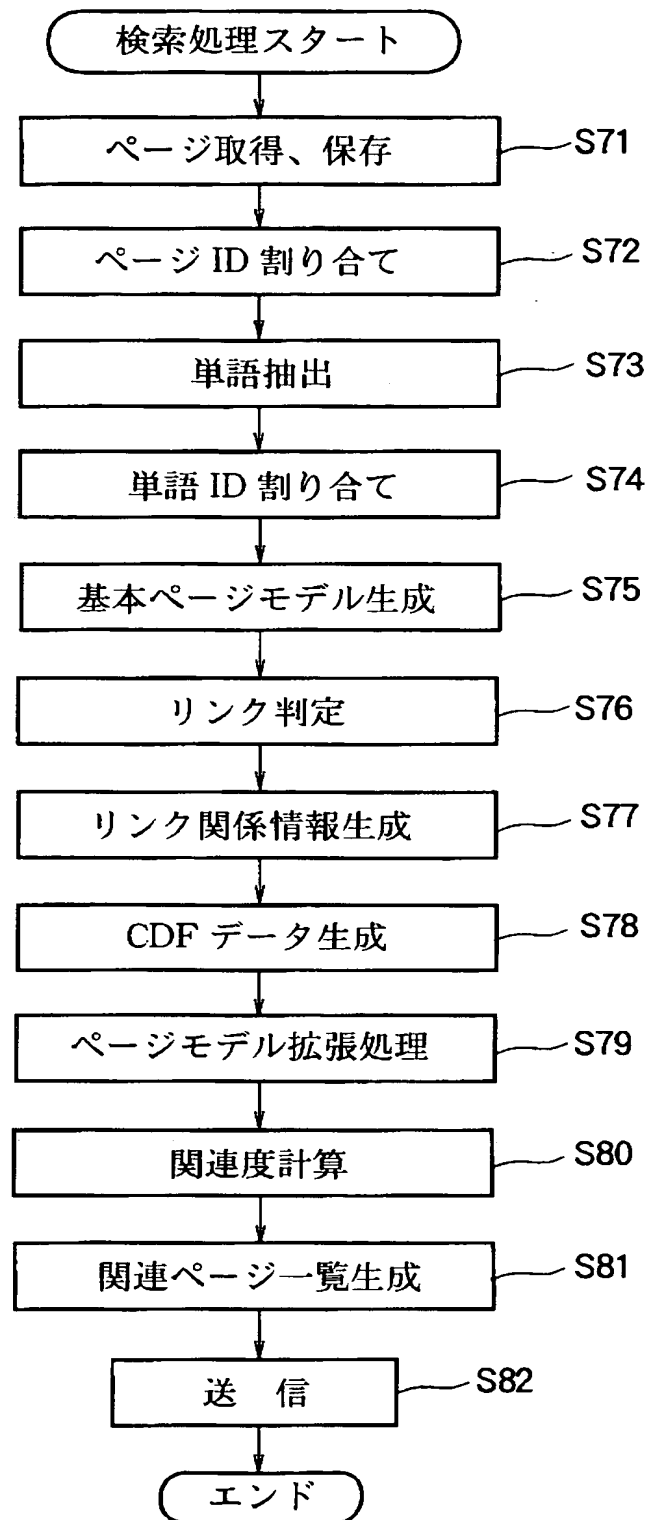


【図 26】



【図 27】

図 27



【図 28】

図 28

ページ ID (Key1)	Co-Parent ページ ID (Key2)	リンクの重み
00010000007	00010000006	1.2
00010000008	00010000006	1.4

261

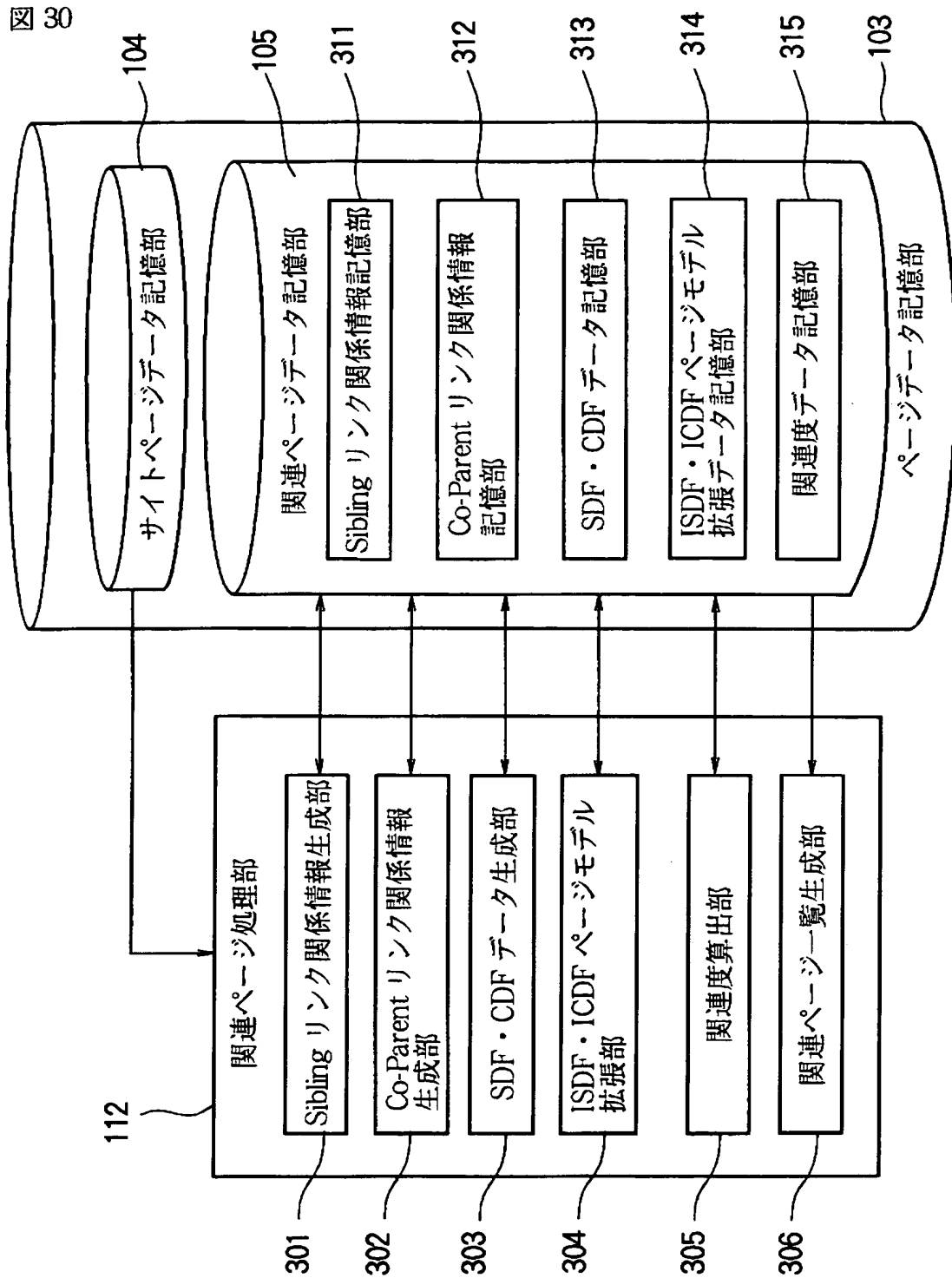
【図 29】

図 29

ページ ID (Key)	ページ ID に含まれる単語 ID と、その 単語 ID を含む Co-Parent ページのリンク の重みの総和
00010000007	{単語 ID, リンクの重みの総和}, {単語 ID, リンクの重みの総和}
00010000008	{単語 ID, リンクの重みの総和}, {単語 ID, リンクの重みの総和}

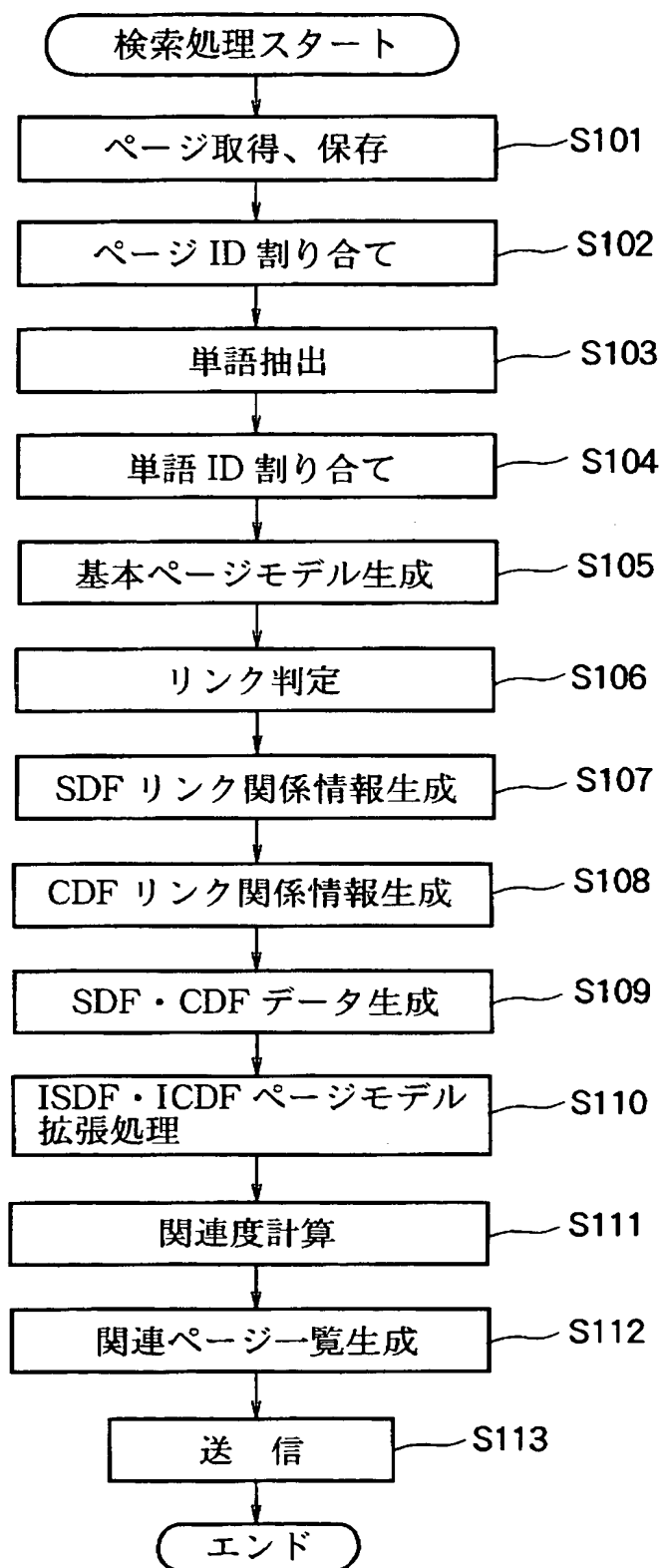
262

【図 30】



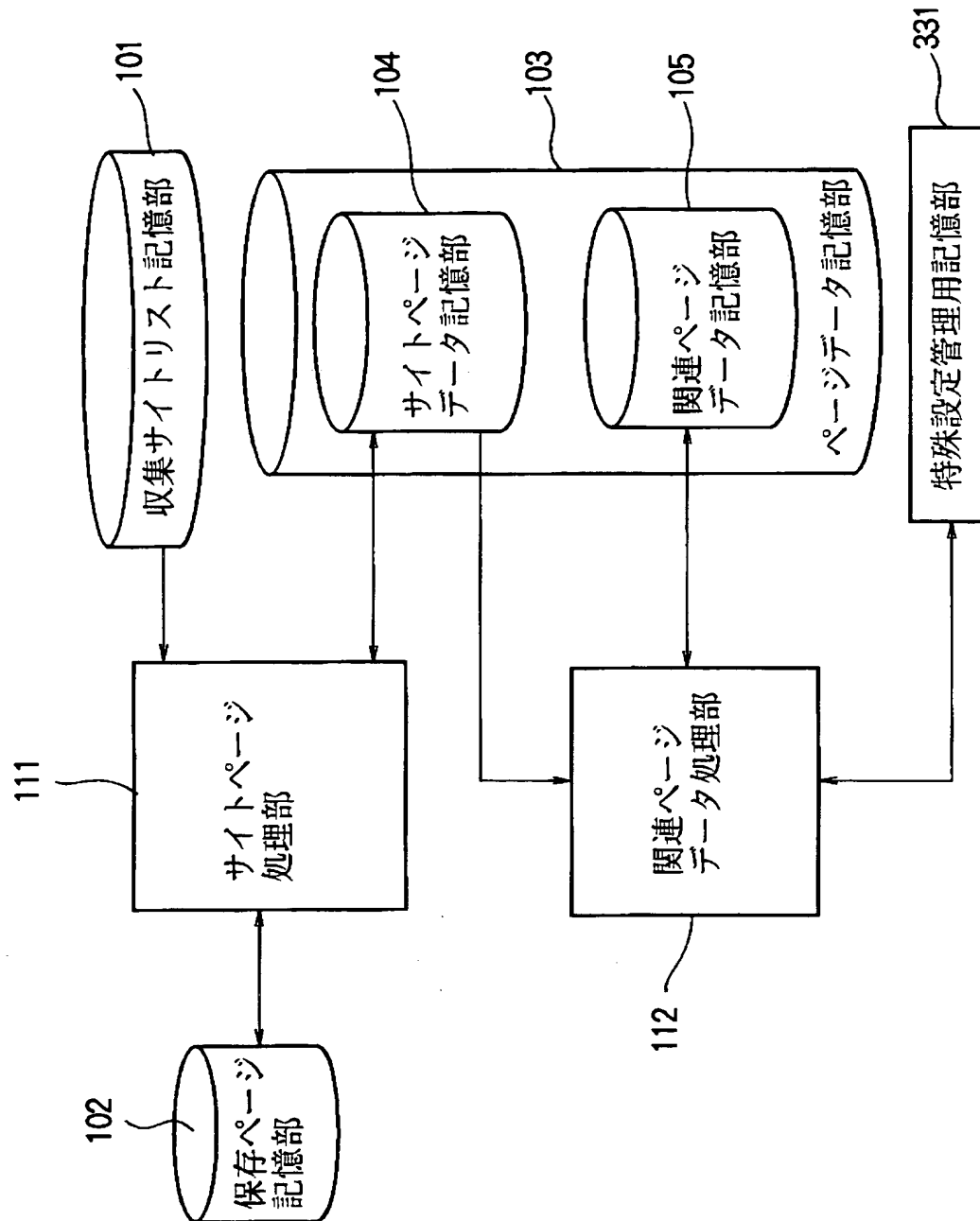
【図 31】

図 31



【図 32】

図 32



【図 3 3】

図 33

タイトル	リンク先 URL (Key)	説明	単語	URL パターン	オーナー ID
ssss News	www.ssss.co.jp/News/ index.html	エスエスエスエス ニュース	最新情報	News	0001
ssss Present	www.ssss.co.jp/Present/ index.html	エスエスエスエス プレゼント	プレゼント	Present	0002

341 特殊設定用管理データ記憶部

【図 3 4】

図 34

オーナー ID (Key)	名前	所属	e-mail	Account	Password
0001	AAAA	HP 室	aaaa@shp.ssss.co.jp	aaaa	aaapwd
0002	SSSS	宣伝	senden@sss.co.jp	senden	sendenpwd

342 特殊設定管理者データ記憶部

【書類名】 要約書

【要約】

【課題】 インターネット上において、ユーザに所定のページの関連ページを精度良く提供できるようにする。

【解決手段】 サイトページ処理部 111 は、サイトに含まれるページを収集し、ページ間の親子関係を判断し、その判断結果をサイトページデータ記憶部 104 に記憶させる。関連ページデータ処理部 112 は、サイトページデータ記憶部 104 に記憶されているデータを用いて、ページ間の兄弟関係と共通親関係の少なくとも一方が考慮された重み付けが施されたページ特徴抽出の値が用いてページ間の関連度を算出する。このページ特徴抽出により、リンク関係にあるページに共通に用いられる単語が関連度算出に大きな影響を与えないように処理される。本発明は、インターネット上に設けられ、所定のページの関連ページ検索するためのサーバに適用することが可能である。

【選択図】 図 5

特願 2 0 0 2 - 3 2 9 4 9 2

出 願 人 履 歴 情 報

識別番号

[0 0 0 0 0 2 1 8 5]

1. 変更年月日

1 9 9 0 年 8 月 3 0 日

[変更理由]

新規登録

住 所

東京都品川区北品川 6 丁目 7 番 3 5 号

氏 名

ソニー株式会社